

# The American Economic Review

THE AMERICAN ECONOMIC REVIEW  
PUBLISHED BY THE AMERICAN ECONOMIC ASSOCIATION  
VOLUME 72, NUMBER 1, FEBRUARY 1982  
PAGES 1-100

# THE AMERICAN ECONOMIC ASSOCIATION

Founded in 1885

• Published at George Banta Co., Inc., Menasha, Wisconsin.

• THE AMERICAN ECONOMIC REVIEW, including four quarterly numbers, the *Proceedings* of the annual meetings, and *Directory* and Supplements, is published by the American Economic Association and is sent to all members five times a year, in March, May, June, September, and December.

• Membership dues of the Association are \$20.00 a year, which includes a year's subscription to both the *American Economic Review* and the *Journal of Economic Literature*. Subscriptions by nonmembers are \$30.00 a year, and only subscriptions to both publications will be accepted. Single copies of the *Review* and *Journal* are \$4.00 each. Each order for copies of either publication must also include a \$.50 per order service charge. Orders should be sent to the Secretary's office, Nashville, Tennessee.

• Correspondence relating to the *Papers and Proceedings*, the *Directory*, advertising, permission to quote, business matters, subscriptions, membership and changes of address may be sent to the secretary, Rendigs Fels, 1313 21st Avenue, South, Nashville, Tennessee 37212. To be effective, notice of change of address must reach the secretary by the 1st of the month previous to the month of publication. The Association's publications are mailed by second class and are not forwardable by the Post Office.

• Second-class postage paid at Nashville, Tennessee and at additional mailing offices. Printed in U.S.A.

## Officers

### *President*

JOHN KENNETH GALBRAITH  
Yale University

### *President-Elect*

KENNETH ARROW  
Harvard University

### *Vice-Presidents*

HENDRIK S. HOUTHAKKER  
Harvard University  
ARTHUR M. OKUN  
Brookings Institution

### *Secretary-Treasurer and Editor of Proceedings*

RENDIGS FELS  
Vanderbilt University

### *Managing Editor of The American Economic Review*

GEORGE H. BORTS  
Brown University

### *Managing Editor of The Journal of Economic Literature*

MARK PERLMAN  
University of Pittsburgh

## Executive Committee

### *Elected Members of the Executive Committee*

ROBERT DORFMAN  
Harvard University  
ARNOLD C. HARBERGER  
University of Chicago  
ROBERT EISNER  
Northwestern University  
JOHN R. MEYER  
Yale University  
GUY HENDERSON ORCUTT  
Yale University  
JOSEPH A. PECHMAN  
Brookings Institution

### *Ex Officio Members*

WASSILY LEONTIEF  
Harvard University  
JAMES TOBIN  
Yale University

Cuk-1102169-65-0293417

# THE AMERICAN ECONOMIC REVIEW

293417

March 1972

VOLUME LXII, NUMBER 1

~~391~~ 391

~~1972~~ '001

VOL. 62 NOS. 1-4

MAR-DEC

1972  
James Tobin

1

## Articles

Inflation and Unemployment

Private and Social Rates of Return to Education  
of Academicians

*Duncan Bailey and Charles Schotta*

19

Distributional Equity and the Optimal Structure  
of Public Prices

*Martin S. Feldstein*

32

The Industrial Composition of U.S. Exports and  
Subsidiary Sales to the Canadian Market

*Thomas Horst*

37

The Process Analysis Alternative to Statistical  
Cost Functions: An Application to Petroleum  
Refining

*James M. Griffin*

46

Optimal Economic Policy and the Problem of In-  
strument Instability

*Robert S. Holbrook*

57

Default Risk, Scale, and the Homemade Leverage  
Theorem

*Vernon L. Smith*

66

The Creation of Risk Aversion by Imperfect  
Capital Markets

*Robert Tempest Masson*

77

Production, Trade, and Protection When There  
are Many Commodities and Two Factors

*William P. Travis*

87

Experimental Evidence on Alternative Portfolio  
Decision Rules

*M. J. Gordon, G. E. Paradis,*

*and C. H. Rorke*

107

Social Returns to Public Information Services:  
Statistical Reporting of U.S. Farm Commodi-  
ties

*Yujiro Hayami and Willis Peterson*

119

GEORGE H. BORTS

Managing Editor

WILMA ST. JOHN

Assistant Editor

## Board of Editors

ROBERT E. BALDWIN

BARBARA R. BERGMANN

JAGDISH N. BHAGWATI

PHILLIP CAGAN

GREGORY C. CHOW

CARL F. CHRIST

JACK HIRSHLEIFER

DANIEL MCFADDEN

ALVIN L. MARTY

EDWIN S. MILLS

HERBERT MOHRING

MARC NERLOVE

EDMUND S. PHELPS

SHERWIN ROSEN

THOMAS R. SAVING

ANNA J. SCHWARTZ

VERNON L. SMITH

• Manuscripts and editorial correspondence relating to the regular quarterly issue of this REVIEW should be addressed to George H. Borts, Managing Editor of THE AMERICAN ECONOMIC REVIEW, Brown University, Providence, R.I. 02912. Manuscripts should be submitted in duplicate and in acceptable form and should be no longer than 50 pages of double-spaced typescript. A \$10 submission fee must accompany each manuscript. *Style Instructions* for guidance in preparing manuscripts will be provided upon request to the editor.

• No responsibility for the views expressed by authors in this REVIEW is assumed by the editors or the publishers, The American Economic Association.

• Copyright American Economic Association 1972.

## Communications

Welfare Economics and Welfare Reform	<i>George Daly and Fred Giertz</i>	131
A "One Line" Proof of the Slutsky Equation	<i>Philip J. Cook</i>	139
A Geometric Treatment of Averch-Johnson's Behavior of the Firm Model:		
Comment	<i>Robert J. Stonebraker</i>	140
Reply	<i>E. E. Zajac</i>	142
Security Pricing and Investment Criteria in Competitive Markets:		
Comment	<i>Prem Kumar</i>	143
Reply	<i>Jan Mossin</i>	147
Macroeconomics of Unbalanced Growth:		
Comment	<i>Michael Keren</i>	149
Reply	<i>William J. Baumol</i>	150
Allais' Restatement of the Quantity Theory of Money: Note	<i>J. L. Scadding</i>	151
The Phillips Curve and the Distribution of Unemployment	<i>A. G. Hines</i>	155
Uncertainty and the Evaluation of Public Investment Decisions:		
Comment	<i>E. J. Mishan</i>	161
Comment	<i>Roland N. McKean and John H. Moore</i>	165
Comment	<i>Alan Nichols</i>	168
Comment	<i>Donald Wellington</i>	170
Reply	<i>Kenneth J. Arrow and Robert C. Lind</i>	171
Optimal Taxes and Pricing:		
Comment	<i>Yew-Kwang Ng</i>	173
Reply	<i>David F. Bradford and William J. Baumol</i>	175
The Role of Money in a Simple Growth Model:		
Comment	<i>Jon Harkness</i>	177
Comment	<i>R. Ramanathan</i>	180
Reply	<i>David Levhari and Don Patinkin</i>	185
A Note on Pollution Prices in a General Equilibrium Model	<i>Larry E. Ruff</i>	186
"Fixed Costs" and the Competitive Firm Under Price Uncertainty:		
Comment	<i>Irwin Bernhardt</i>	193
Reply	<i>Agnar Sandmo</i>	194
Separability and Complementarity	<i>Eugene Silberberg</i>	196
Money Illusion and the Aggregate Consumption Function:		
Comment	<i>Alex Cukierman</i>	198
Reply	<i>William H. Branson and Alvin K. Klevorick</i>	207
Do Blacks Save More?	<i>Marjorie Galenson</i>	211
On Measuring the Nearness of Near-Moneys:		
Comment	<i>Tong Hun Lee</i>	217
Comment	<i>Larry Steinhauer and John Chang</i>	221
Reply	<i>V. K. Chetty</i>	226
Lags in the Effects of Monetary Policy:		
Comment	<i>Paul E. Smith</i>	230
Reply and Some Further Thoughts	<i>J. Ernest Tanner</i>	234
Errata		238
Announcement		239
Notes		240

# THE AMERICAN ECONOMIC REVIEW

GEORGE H. BORTS  
Managing Editor

WILMA ST. JOHN  
Assistant Editor

## Board of Editors

ROBERT E. BALDWIN  
BARBARA R. BERGMANN  
JAGDISH N. BHAGWATI  
PHILLIP CAGAN  
GREGORY C. CHOW  
CARL F. CHRIST  
JACK HIRSHLEIFER  
DANIEL MCFADDEN  
ALVIN L. MARTY  
EDWIN S. MILLS  
HERBERT MOHRING  
MARC NERLOVE  
EDMUND S. PHELPS  
SHERWIN ROSEN  
THOMAS R. SAVING  
ANNA J. SCHWARTZ  
VERNON L. SMITH

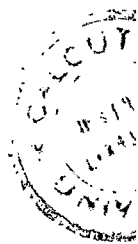
• Manuscripts and editorial correspondence relating to the regular quarterly issue of this REVIEW should be addressed to George H. Borts, Managing Editor of THE AMERICAN ECONOMIC REVIEW, Brown University, Providence, R.I. 02912. Manuscripts should be submitted in duplicate and in acceptable form and should be no longer than 50 pages of double-spaced typescript. A \$10 submission fee must accompany each manuscript. *Style Instructions* for guidance in preparing manuscripts will be provided upon request to the editor.

• No responsibility for the views expressed by authors in this REVIEW is assumed by the editors or the publishers, The American Economic Association.

• Copyright American Economic Association 1972.

June 1972

VOLUME LXII, NUMBER 3



## Articles

Maximum Principles in Analytical Economics  
*Paul A. Samuelson* 249

Housing Market Discrimination, Homeownership,  
and Savings Behavior  
*John F. Kain and John M. Quigley* 263

Theory of the Firm Facing Uncertain Demand  
*Hayne E. Leland* 278

The Administered-Price Thesis Reconfirmed  
*Gardiner C. Means* 292

On Taxation and the Control of Externalities  
*William J. Baumol* 307

The Effects of Minimum Wages on the Distribu-  
tion of Changes in Aggregate Employment  
*Marvin Kesters and Finis Welch* 323

Implications of the Theory of Rationing for Con-  
sumer Choice Under Uncertainty  
*Peter A. Diamond and Menahem Yaari* 333

Life Cycle Saving: Theory and Fact  
*Keizo Nagatani* 344

The Rationale of the Mean-Standard Deviation  
Analysis, Skewness Preference, and the Demand  
for Money  
*S. C. Tsiang* 354

Black Education, Earnings, and Interregional  
Migration: Some New Evidence.  
*Leonard Weiss and Jeffrey G. Williamson* 372

Incentive Contracts and Competitive Bidding  
*David P. Baron* 384

## Communications

Price Discrimination by Regulated Motor Carriers	<i>Josephine E. Olson</i>	395
The Dynamics of Firm Behavior Under Alternative Cost Structures	<i>George A. Hay</i>	403
Urban Poverty and Labor Force Participation: Note	<i>Larry Sawers</i>	414
Nordhaus' Theory of Optimal Patent Life:		
A Geometric Reinterpretation	<i>F. M. Scherer</i>	422
The Optimum Life of a Patent: Reply	<i>William D. Nordhaus</i>	428
A Simple Approach to Existence and Uniqueness of Competitive Equilibria	<i>Donald W. Katzner</i>	432
Stochastic Dominance vs. Mean-Variance Portfolio Analysis: An Empirical Evaluation	<i>R. Burr Porter and Jack E. Gaumnitz</i>	438
Progression and Leisure	<i>M. G. Allingham</i>	447
Peasants, Procreation, and Pensions:		
Comment	<i>Marianne Abeles Ferber</i>	451
Reply	<i>Philip Neher</i>	452
Upward Sloping Demand Curves Without the Giffen Paradox	<i>Daniel C. Vandermeulen</i>	453
Determinants of the Commodity Structure of U.S. Trade:		
Comment:	<i>Lawrence Weiser and Keith Jay</i>	459
Reply	<i>Robert E. Baldwin</i>	465
Unions and Relative Real Wages	<i>Michael J. Boskin</i>	466
Commodity Price Equalization: A Note on Factor Mobility and Trade	<i>Frank Flatters</i>	473
A Theory and Test of Credit Rationing:		
Some Generalizations	<i>Vernon L. Smith</i>	477
Further Notes	<i>Dwight M. Jaffee</i>	484
Distributional Equality and Aggregate Utility:		
Further Comment	<i>Maurice McManus, Gary M. Walton, and Richard B. Coffman</i>	489
Further Comment	<i>Roger A. McCain</i>	497
Reply	<i>William Breit and William Patton Culbertson, Jr.</i>	501
Announcement		503
Notes		504

# THE AMERICAN ECONOMIC REVIEW

GEORGE H. BORTS

Managing Editor

WILMA ST. JOHN

Assistant Editor

## Board of Editors

ROBERT E. BALDWIN  
BARBARA R. BERGMANN  
JAGDISH N. BHAGWATI  
PHILLIP CAGAN  
GREGORY C. CHOW  
CARL F. CHRIST  
JACK HIRSHLEIFER  
DANIEL MCFADDEN  
ALVIN L. MARTY  
EDWIN S. MILLS  
HERBERT MOHRING  
MARC NERLOVE  
EDMUND S. PHELPS  
SHERWIN ROSEN  
THOMAS R. SAVING  
ANNA J. SCHWARTZ  
VERNON L. SMITH

• Manuscripts and editorial correspondence relating to the regular quarterly issue of this REVIEW should be addressed to George H. Borts, Managing Editor of THE AMERICAN ECONOMIC REVIEW, Brown University, Providence, R.I. 02912. Manuscripts should be submitted in duplicate and in acceptable form and should be no longer than 50 pages of double-spaced typescript. A \$10 submission fee must accompany each manuscript. *Style Instructions* for guidance in preparing manuscripts will be provided upon request to the editor.

• No responsibility for the views expressed by authors in this REVIEW is assumed by the editors or the publishers, The American Economic Association.

• Copyright American Economic Association 1972.

September 1972

VOLUME LXII, NUMBER 4



## Articles

- The 1972 Report of the President's Council of Economic Advisers: Inflation and Unemployment *Edgar L. Feige* 509
- The 1972 Report of the President's Council of Economic Advisers: International Aspects *Peter B. Kenen* 517
- The 1972 Report of the President's Council of Economic Advisers: Inflation and Controls *Reuben A. Kessel* 527
- The 1972 Report of the President's Council of Economic Advisers: Economics and Government *Edmund S. Phelps* 533
- Money, Income, and Causality *Christopher A. Sims* 540
- Neoclassical Investment Models and French Private Manufacturing Investment *Richard Schramm* 553
- Anticipatory and Objective Models of Durable Goods Demand *F. Thomas Juster and Paul Wachtel* 564
- Disasters and Charity: Some Aspects of Cooperative Economic Behavior *Christopher M. Douty* 580
- Optimization and Scale Economies in Urban Bus Transportation *Herbert Mohring* 591
- The Economics of Environmental Preservation: A Theoretical and Empirical Analysis *Anthony C. Fisher, John V. Krutilla, and Charles J. Cicchetti* 605
- The Preventive Tariff and the Dual in Linear Programming *Leonard Waverman* 620
- A Choice-Theoretic Model of an Income-Investment Accelerator *Herschel I. Grossman* 630
- Advertising and the Aggregate Consumption Function *Lester D. Taylor and Daniel Weiserbs* 642

## Communications

A Note on the Stigler-Kindahl Study of Industrial Prices	<i>George A. Hay</i>	656
The Statistical Theory of Racism and Sexism	<i>Edmund S. Phelps</i>	659
Learning and Productivity Change in Metal Products	<i>Leonard Dudley</i>	662
The Number of Firms and Competition	<i>Eugene F. Fama and Arthur B. Laffer</i>	670
Soviet Postwar Economic Growth and Capital-Labor Substitution:		
Comment	<i>Earl R. Brubaker</i>	675
Comment	<i>Mitchell Kellman and Lorenzo L. Perez</i>	679
Reply	<i>Martin L. Weitzman</i>	682
A Model of Soviet-Type Economic Planning:		
Comment	<i>J. M. Montias</i>	685
Reply	<i>Michael Manove</i>	689
The Determinants of U.S. Direct Investment in the E.E.C.:		
Comment	<i>Murray A. Goldberg</i>	692
Reply	<i>Anthony E. Scaperlanda and Laurence J. Mauer</i>	700
Monopoly Output Under Alternative Spatial Pricing Techniques	<i>M. L. Greenhut and H. Ohta</i>	705
Job Search, the Duration of Unemployment, and the Phillips Curve:		
Comment	<i>Paul Gayer and Robert S. Goldfarb</i>	714
Reply	<i>Dale T. Mortensen</i>	718
Production Indeterminacy with Three Goods and Two Factors:		
Rejoinder	<i>Douglas B. Stewart</i>	720
The Last Word?	<i>James R. Melvin</i>	723
An Analysis of Turning Point Forecasts	<i>H. O. Stekler</i>	724
The Permanent Income Hypothesis: Evidence from Time-Series Data	<i>Prem S. Laumas and Khan A. Mohabbat</i>	730
The Incidence of the Social Security Payroll Tax:		
Comment	<i>Martin S. Feldstein</i>	735
Reply	<i>John A. Brittain</i>	739
Decision Rules for Effective Protection in Less Developed Economies	<i>Trent J. Bertrand</i>	743
Substitution, Complementarity, and the Residual Variation: Some Further Results	<i>Louis Phelps and Philippe Rouzier</i>	747
Schooling and Earnings of Low Achievers:		
Comment	<i>Barry R. Chiswick</i>	752
Comment	<i>Stanley Masters and Thomas Ribich</i>	755
Reply	<i>W. Lee Hansen, Burton A. Weisbrod,     and William J. Scanlon</i>	760
Errata		763
Statement of Editorial Policy		764
Notes		765
Announcements		772

# THE AMERICAN ECONOMIC REVIEW

GEORGE H. BORTS  
Managing Editor

WILMA ST. JOHN  
Assistant Editor

## Board of Editors

ROBERT E. BALDWIN  
BARBARA R. BERGMANN  
JAGDISH N. BHAGWATI  
PHILLIP CAGAN  
GREGORY C. CHOW  
CARL F. CHRIST  
JACK HIRSHLEIFER  
DANIEL MCFADDEN  
ALVIN L. MARTY  
EDWIN S. MILLS  
HERBERT MOHRING  
MARC NERLOVE  
EDMUND S. PHELPS  
SHERWIN ROSEN  
THOMAS R. SAVING  
ANNA J. SCHWARTZ  
VERNON L. SMITH

• Manuscripts and editorial correspondence relating to the regular quarterly issue of this REVIEW should be addressed to George H. Borts, Managing Editor of THE AMERICAN ECONOMIC REVIEW, Brown University, Providence, R.I. 02912. Manuscripts should be submitted in duplicate and in acceptable form and should be no longer than 50 pages of double-spaced typescript. A \$10 submission fee must accompany each manuscript. *Style Instructions* for guidance in preparing manuscripts will be provided upon request to the editor.

• No responsibility for the views expressed by authors in this REVIEW is assumed by the editors or the publishers, The American Economic Association.

• Copyright American Economic Association 1972.

December 1972

VOLUME LXII, NUMBER 5



## Articles

Production, Information Costs, and Economic Organization

*Armen A. Alchian and Harold Demsetz* 777

Education and Underemployment in the Urban Ghetto

*Bennett Harrison* 796

Option Demand and Consumer's Surplus: Valuing Price Changes under Uncertainty

*Richard Schmalensee* 813

Sectoral Investment Determination in a Developing Economy

*Jere R. Behrman* 825

Capital Deepening Response in an Economy with Heterogeneous Capital Goods

*Edwin Burmeister and Stephen J. Turnovsky* 842

Interest Rates and Inflationary Expectations: New Evidence

*William E. Gibson* 854

Capital Gains and the Aggregate Consumption Function

*Kul B. Bhatia* 866

Keynes-Wicksell and Neoclassical Models of Money and Growth

*Stanley Fischer* 880

The Demand for the Services of Non-Federal Governments

*Thomas E. Borchering and Robert T. Deacon* 891

The Prediction Performance of the FRB-MIT-PENN Model of the U.S. Economy

*Charles R. Nelson* 902

Uncertain Entry and Excess Capacity

*Morton I. Kamien and Nancy L. Schwartz* 918

The Allocation of Transitory Income Among Consumers' Assets

*Michael R. Darby* 928

## Communications

Duality and the Many Consumer's Surpluses	<i>Eugene Silberberg</i>	942
Fiscal and Monetary Policy Reconsidered:		
Comment	<i>Jack Vernon</i>	953
Further Reply	<i>Robert Eisner</i>	957
Pollution and Pricing	<i>Allen V. Kneese</i>	958
Local versus National Pollution Control: Note	<i>Sam Peltzman and T. Nicolaus Tideman</i>	959
Behavior of the Firm Under Regulatory Constraint	<i>Jerome L. Stein and George H. Borts</i>	964
The Futility of Pareto-Efficient Distributions	<i>E. J. Mishan</i>	971
Peasants, Procreation, and Pensions:		
Comment	<i>Warren C. Robinson</i>	977
Reply	<i>Philip A. Neher</i>	979
A New Look at the Muth Model	<i>Wyatt Mankin</i>	980
Choice Involving Unwanted Risky Events and Optimal Insurance	<i>J. M. Parkin and S. Y. Wu</i>	982
A Spectral Analysis of Post-Accord Federal Open Market Operations:		
Comment	<i>Llad Phillips and Robert Weintraub</i>	988
Reply	<i>Vittorio Bonomo and Charles Schotta</i>	993
Multi-Neutral Technical Progress: Compatibilities, Conditions, and Consistency with		
Some Evidence	<i>Earl R. Brubaker</i>	997
Inflation, Unemployment, and Economic Welfare:		
Comment	<i>Gordon Tullock</i>	1004
Reply	<i>Roger N. Waud</i>	1005
Lerner on Pollution Controls:		
Comment	<i>Ali M. Reza</i>	1007
Pollution Abatement Subsidies	<i>Abba P. Lerner</i>	1009
Price-Quantity Adjustments in a Competitive Market	<i>E. C. H. Veendorp</i>	1011
Notes		1016
Proposed Washington Newsletter		1023
Titles of Doctoral Dissertations		1029

# THE AMERICAN ECONOMIC REVIEW

VOLUME LXII

## BOARD OF EDITORS

ROBERT E. BALDWIN	EDWIN S. MILLS
BARBARA R. BERGMANN	HERBERT MOHRING
JAGDISH N. BHAGWATI	MARC NERLOVE
PHILLIP CAGAN	EDMUND S. PHELPS
GREGORY C. CHOW	SHERWIN ROSEN
CARL F. CHRIST	THOMAS R. SAVING
JACK HIRSHLEIFER	ANNA J. SCHWARTZ
DANIEL MCFADDEN	VERNON L. SMITH
ALVIN L. MARTY	

MANAGING EDITOR  
GEORGE H. BORTS

THE AMERICAN ECONOMIC ASSOCIATION

Executive Office: Nashville, Tennessee

Editorial Office: Brown University, Providence, Rhode Island

Copyright 1972  
AMERICAN ECONOMIC ASSOCIATION

# CONTENTS OF ARTICLES AND COMMUNICATIONS

J. Tobin: Inflation and Unemployment.....	1	Y-K Ng: Optimal Taxes and Pricing: Comment.....	173
D. Bailey and C. Schotta: Private and Social Rates of Return to Education of Academicians.....	19	D. F. Bradford and W. J. Baumol: Reply....	175
M. S. Feldstein: Distributional Equity and the Optimal Structure of Public Prices.....	32	J. Harkness: The Role of Money in a Simple Growth Model: Comment.....	177
T. Horst: The Industrial Composition of U.S. Exports and Subsidiary Sales to the Canadian Market.....	37	R. Ramanathan: Comment.....	180
J. M. Griffin: The Process Analysis Alternative to Statistical Cost Functions: An Application to Petroleum Refining.....	46	D. Levhari and D. Patinkin: Reply.....	185
R. S. Holbrook: Optimal Economic Policy and the Problem of Instrument Instability.....	57	L. E. Ruff: A Note on Pollution Prices in a General Equilibrium Model.....	186
V. L. Smith: Default Risk, Scale, and the Homemade Leverage Theorem.....	66	I. Bernhardt: "Fixed Costs" and the Competitive Firm Under Price Uncertainty: Comment.....	193
R. T. Masson: The Creation of Risk Aversion by Imperfect Capital Markets.....	77	A. Sandmo: Reply.....	194
W. R. Travis: Production, Trade, and Protection When There are Many Commodities and Two Factors.....	87	E. Silberberg: Separability and Complementarity.....	196
M. J. Gordon, G. E. Paradis, and C. H. Rorke: Experimental Evidence on Alternative Portfolio Decision Rules.....	107	A. Cukierman: Money Illusion and the Aggregate Consumption Function: Comment..	198
Y. Hayami and W. Peterson: Social Returns to Public Information Services: Statistical Reporting of U. S. Farm Commodities.....	119	W. H. Branson and A. K. Klevorick: Reply....	207
G. Daly and F. Giertz: Welfare Economics and Welfare Reform.....	131	M. Galenson: Do Blacks Save More?.....	211
P. J. Cook: A "One Line" Proof of the Slutsky Equation.....	139	T. H. Lee: On Measuring the Nearness of Near-Moneys: Comment.....	217
R. J. Stonebraker: A Geometric Treatment of Averch-Johnson's Behavior of the Firm Model: Comment.....	140	L. Steinhauer and J. Chang: Comment.....	221
E. E. Zajac: Reply.....	142	V. K. Chetty: Reply.....	226
P. Kumar: Security Pricing and Investment Criteria in Competitive Markets: Comment.	143	P. E. Smith: Lags in the Effects of Monetary Policy: Comment.....	230
J. Mossin: Reply.....	147	J. E. Tanner: Reply and Some Further Thoughts.....	234
M. Keren: Macroeconomics of Unbalanced Growth: Comment.....	149	P. A. Samuelson: Maximum Principles in Analytical Economics.....	249
W. J. Baumol: Reply.....	150	J. F. Kain and J. M. Quigley: Housing Market Discrimination, Homeownership, and Savings Behavior.....	263
J. L. Scadding: Allais' Restatement of the Quantity Theory of Money: Note.....	151	H. E. Leland: Theory of the Firm Facing Uncertain Demand.....	278
A. G. Hines: The Phillips Curve and the Distribution of Unemployment.....	155	G. C. Means: The Administered-Price Thesis Reconfirmed.....	292
E. J. Mishan: Uncertainty and the Evaluation of Public Investment Decisions: Comment..	161	W. J. Baumol: On Taxation and the Control of Externalities.....	307
R. N. McKean and J. H. Moore: Comment...	165	M. Kusters and F. Welch: The Effects of Minimum Wages on the Distribution of Changes in Aggregate Employment.....	323
A. Nichols: Comment.....	168	P. A. Diamond and M. Yaari: Implications of the Theory of Rationing for Consumer Choice Under Uncertainty.....	333
D. Wellington: Comment.....	170	K. Nagatani: Life Cycle Saving: Theory and Fact.....	344
K. J. Arrow and R. C. Lind: Reply.....	171	S. C. Tsiang: The Rationale of the Mean-Standard Deviation Analysis, Skewness Preference, and the Demand for Money.....	354
		L. Weiss and J. G. Williamson: Black Educa-	

tion, Earnings, and Interregional Migration: Some New Evidence.....	372	nomics and Government.....	533
D. P. Baron: Incentive Contracts and Competitive Bidding.....	384	C. A. Sims: Money, Income, and Causality....	540
J. E. Olson: Price Discrimination by Regulated Motor Carriers.....	395	R. Schramm: Neoclassical Investment Models and French Private Manufacturing Investment.....	553
G. A. Hay: The Dynamics of Firm Behavior Under Alternative Cost Structures.....	403	F. T. Juster and P. Wachtel: Anticipatory and Objective Models of Durable Goods Demand.....	564
L. Sawers: Urban Poverty and Labor Force Participation: Note.....	414	C. M. Douy: Disasters and Charity: Some Aspects of Co-operative Economic Behavior.....	580
F. M. Scherer: Nordhaus' Theory of Optimal Patent Life: A Geometric Reinterpretation.....	422	H. Mohring: Optimization and Scale Economies in Urban Bus Transportation.....	591
W. D. Nordhuas: The Optimum Life of a Patent: Reply.....	428	A. C. Fisher, J. V. Krutilla, and C. J. Cicchetti: The Economics of Environmental Preservation: A Theoretical and Empirical Analysis.....	605
D. W. Katzner: A Simple Approach to Existence and Uniqueness of Competitive Equilibria.....	432	L. Waverman: The Preventive Tariff and the Dual in Linear Programming.....	620
R. B. Porter and J. E. Gaumnitz: Stochastic Dominance vs. Mean-Variance Portfolio Analysis: An Empirical Evaluation.....	438	H. I. Grossman: A Choice-Theoretic Model of an Income-Investment Accelerator.....	630
M. G. Allingham: Progression and Leisure....	447	L. D. Taylor and D. Weiserbs: Advertising and the Aggregate Consumption Function....	642
M. A. Ferber: Peasants, Procreation, and Pensions: Comment.....	451	G. A. Hay: A Note on the Stigler-Kindahl Study of Industrial Prices.....	656
P. Neher: Reply.....	452	E. S. Phelps: The Statistical Theory of Racism and Sexism.....	659
D. C. Vandermeulen: Upward Sloping Demand Curves Without the Giffen Paradox..	453	L. Dudley: Learning and Productivity Change in Metal Products.....	662
L. Weiser and K. Jay: Determinants of the Commodity Structure of U.S. Trade: Comment.....	459	E. F. Fama and A. B. Laffer: The Number of Firms and Competition.....	670
R. E. Baldwin: Reply.....	465	E. R. Brubaker: Soviet Postwar Economic Growth and Capital-Labor Substitution: Comment.....	675
M. J. Boskin: Unions and Relative Real Wages.....	466	M. Kellman and L. L. Perez: Comment.....	679
F. Flatters: Commodity Price Equilization: A Note on Factor Mobility and Trade.....	473	M. L. Weitzman: Reply.....	682
V. L. Smith: A Theory and Test of Credit Rationing: Some Generalizations.....	477	J. M. Montias: A Model of Soviet-Type Economic Planning: Comment.....	685
D. M. Jaffee: Further Notes.....	484	M. Manove: Reply.....	689
M. McManus, G. M. Walton, and R. B. Coffman: Distributional Equality and Aggregate Utility: Further Comment.....	489	M. A. Goldberg: The Determinants of U.S. Direct Investment in the E.E.C.: Comment....	692
R. A. McCain: Further Comment.....	497	A. E. Scaperlanda and L. J. Mauer: Reply....	700
W. Breit and W. P. Culbertson, Jr.: Reply....	501	M. L. Greenhut and H. Ohta: Monopoly Output Under Alternative Spatial Pricing Techniques.....	705
E. L. Feige: The 1972 Report of the President's Council of Economic Advisers: Inflation and Unemployment.....	509	P. Gayer and R. S. Goldfarb: Job Search, the Duration of Unemployment, and the Phillips Curve: Comment.....	714
P. B. Kenen: The 1972 Report of the President's Council of Economic Advisers: International Aspects.....	517	D. T. Mortensen: Reply.....	718
R. A. Kessel: The 1972 Report of the President's Council of Economic Advisers: Inflation and Controls.....	527	D. B. Stewart: Production Indeterminacy with Three Goods and Two Factors: Rejoinder.....	720
E. S. Phelps: The 1972 Report of the President's Council of Economic Advisers: Eco-		J. R. Melvin: The Last Word?.....	723
		H. O. Stekler: An Analysis of Turning Point Forecasts.....	724

<b>P. S. Laumas and K. A. Mohabbat:</b> The Permanent Income Hypothesis: Evidence from Time-Series Data.....	730	<b>M. I. Kamien and N. L. Schwartz:</b> Uncertain Entry and Excess Capacity.....	918
<b>M. S. Feldstein:</b> The Incidence of the Social Security Payroll Tax: Comment.....	735	<b>M. R. Darby:</b> The Allocation of Transitory Income Among Consumers' Assets.....	928
<b>J. A. Brittain:</b> Reply.....	739	<b>E. Silberberg:</b> Duality and the Many Consumer's Surpluses.....	942
<b>T. J. Bertrand:</b> Decision Rules for Effective Protection in Less Developed Economies....	743	<b>J. Vernon:</b> Fiscal and Monetary Policy Reconsidered: Comment.....	953
<b>L. Phlips and P. Rouzier:</b> Substitution, Complementarity, and the Residual Variation: Some Further Results.....	747	<b>R. Eisner:</b> Further Reply.....	957
<b>B. R. Chiswick:</b> Schooling and Earnings of Low Achievers: Comment.....	752	<b>A. V. Kneese:</b> Pollution and Pricing.....	958
<b>S. Masters and T. Ribich:</b> Comment.....	755	<b>S. Peltzman and T. N. Tideman:</b> Local versus National Pollution Control: Note.....	959
<b>W. L. Hansen, B. A. Weisbrod, and W. J. Scanlon:</b> Reply.....	760	<b>J. L. Stein and G. H. Borts:</b> Behavior of the Firm Under Regulatory Constraint.....	964
<b>A. A. Alchian and H. Demsetz:</b> Production, Information Costs, and Economic Organization.....	777	<b>E. J. Mishan:</b> The Futility of Pareto-Efficient Distributions.....	971
<b>B. Harrison:</b> Education and Underemployment in the Urban Ghetto.....	796	<b>W. C. Robinson:</b> Peasants, Procreation, and Pensions: Comment.....	977
<b>R. Schmalensee:</b> Option Demand and Consumer's Surplus: Valuing Price Changes Under Uncertainty.....	813	<b>P. A. Neher:</b> Reply.....	979
<b>J. R. Behrman:</b> Sectoral Investment Determination in a Developing Economy.....	825	<b>W. Mankin:</b> A New Look at the Muth Model.....	980
<b>E. Burmeister and S. J. Turnovsky:</b> Capital Deepening Response in an Economy with Heterogeneous Capital Goods.....	842	<b>J. M. Parkin and S. Y. Wu:</b> Choice Involving Unwanted Risky Events and Optimal Insurance.....	982
<b>W. E. Gibson:</b> Interest Rates and Inflationary Expectations: New Evidence.....	854	<b>L. Phillips and R. Weintraub:</b> A Spectral Analysis of Post-Accord Federal Open Market Operations: Comment.....	988
<b>K. B. Bhatia:</b> Capital Gains and the Aggregate Consumption Function.....	866	<b>V. Bonomo and C. Schotta:</b> Reply.....	993
<b>S. Fischer:</b> Keynes-Wicksell and Neoclassical Models of Money and Growth.....	880	<b>E. R. Brubaker:</b> Multi-Neutral Technical Progress: Compatibilities, Conditions, and Consistency with Some Evidence.....	997
<b>T. E. Borcharding and R. T. Deacon:</b> The Demand for the Services on Non-Federal Governments.....	891	<b>G. Tullock:</b> Inflation, Unemployment, and Economic Welfare: Comment.....	1004
<b>C. R. Nelson:</b> The Prediction Performance of the FRB-MIT-PENN Model of the U.S. Economy.....	902	<b>R. N. Waud:</b> Reply.....	1005
		<b>A. M. Reza:</b> Lerner on Pollution Controls: Comment.....	1007
		<b>A. P. Lerner:</b> Pollution Abatement Subsidies..	1009
		<b>E. C. H. Veendorp:</b> Price-Quantity Adjustments in a Competitive Market.....	1011



# CONTENTS OF PAPERS AND PROCEEDINGS

## *Richard T. Ely Lecture*

- J. Robinson:** The Second Crisis of Economic Theory..... 1

## *Have Fiscal and/or Monetary Policies Failed?*

- M. Friedman:** Have Monetary Policies Failed?..... 10

- J. G. Gurley:** Have Fiscal and Monetary Policies Failed?..... 19

- A. M. Okun:** Have Fiscal and/or Monetary Policies Failed?..... 24

## *Issues and Perspectives of Black Political Economy*

- W. K. Tabb:** Viewing Minority Economic Development as a Problem in Political Economy..... 31

- R. S. Browne:** The Economic Case for Reparations to Black America..... 39

## *Regulatory Reform in Transportation*

- G. W. Hilton:** The Basic Behavior of Regulatory Commissions..... 47

- R. J. Sampson:** Inherent Advantages Under Regulation..... 55

## *Arts in the Affluent Society*

- T. Scitovsky:** What's Wrong with the Arts is What's Wrong with Society..... 62

- H. R. Faine:** Unions and the Arts..... 70

## *Models for Urban Land Use, Housing, and Transportation*

- L. Merewitz:** Public Transportation: Wish Fulfillment and Reality in the San Francisco Bay Area..... 78

- R. F. Engle III, F. M. Fisher, J. R. Harris, and J. Rothenberg:** An Econometric Simulation Model of Intra-Metropolitan Housing Location: Housing, Business, Transportation, and Local Government..... 87

- Discussion by **R. F. Muth** and **D. F. Bradford**.. 98

## *The Corporation, Technology, and the State*

- R. Marris:** Is the Corporate Economy a Corporate State?..... 103

## *The Keynesian Revolution and Its Pioneers*

- A. R. Sweezy:** The Keynesians and Government Policy, 1933-1939..... 116

- B. R. Jones:** The Role of Keynesians in War-time Policy and Postwar Planning, 1940-1946..... 125

- Discussion by **L. Keyserling**, **R. R. Nathan**, and **L. B. Currie**..... 134

## *Economists and the Trade Union Movement*

- R. Lekachman:** Academic Wisdom and Union Reality..... 142

- B. Sexton:** The Working Class Experience... 149

- Discussion by **M. Strober**..... 154

## *What Economic Equality for Women Requires*

- H. Zellner:** Discrimination Against Women, Occupational Segregation, and the Relative Wage..... 157

- F. B. Weisskoff:** "Women's Place" in the Labor Market..... 161

- C. D. Phelps:** Is the Household Obsolete?.... 167

- Discussion by **S. H. Sandell**..... 175

## *Some Contradictions of Capitalism*

- H. M. Cleaver, Jr.:** The Contradictions of the Green Revolution..... 177

- H. Wachtel:** Capitalism and Poverty in America: Paradox or Contradiction?..... 187

## *Is the South Still Backward?*

- M. I. Foster:** Is the South Still a Backward Region, and Why?..... 195

- F. Ray Marshall:** Some Rural Economic Developments in the South..... 204

## *The Stock Market and the Economy*

- I. Friend:** The Economic Consequences of the Stock Market..... 212

- R. Rasche:** Impact of the Stock Market on Private Demand..... 220

- Discussion by **F. Modigliani**..... 229

## *Issues in Incomes Policy*

- G. Haberler:** Incomes Policy and Inflation: Some Further Reflections..... 234

- L. Ulman:** Cost-Push and Some Policy Alternatives..... 242

- A. R. Weber:** A Wage-Price Freeze as an Instrument of Incomes Policy: Or the Blizzard of '71..... 251

## *The Future of Consumer Sovereignty*

- A. P. Lerner:** The Economics and Politics of Consumer Sovereignty..... 258

- H. Gintis:** Consumer Behavior and the Concept of Sovereignty: Explanations of Social Decay..... 267

## *The Economics of the Military-Industrial Complex*

- W. Adams** and **W. J. Adams:** The Military-In-

dustrial Complex: A Market Structure Analysis.....	279	L. S. Silk: Truth vs. Partisan Political Purpose.....	376
R. Kaufman: MIRVing the Boondoggle: Contracts, Subsidy, and Welfare in the Aerospace Industry.....	288	J. R. Meyer: Communications Gap Narrows but Persists.....	379
M. Reich: Does the U.S. Economy Require Military Spending?.....	296	B. D. Nossiter: Economists and Reporters: A Combination to Promote/Restrain Trade....	381
J. R. Kurth: The Political Economy of Weapons Procurement: The Follow-On Imperative.....	304	H. C. Wallich: Economists and the Press—A Progress Report.....	384
S. Melman: Ten Propositions on the War Economy.....	312	R. M. Janssen: Friends with Points of Friction and Misunderstanding.....	387
		J. A. Schnittker: Economists and Economic Reporters.....	389
<i>Taxation of the Poor and the Rich</i>		<i>On the Emerging Problems of Development Policy</i>	
D. M. Gordon: Taxation of the Poor and the Normative Theory of Tax Incidence.....	319	A. Fishlow: Brazilian Size Distribution of Income.....	391
E. A. Thompson: The Taxation of Wealth and the Wealthy.....	329	R. A. Berry: Farm Size Distribution, Income Distribution, and the Efficiency of Agricultural Production: Colombia.....	403
Discussion by R. J. Lampman and H. J. Aaron.....	331		
<i>On the Status and Relevance of Economic Theory</i>		<i>History of Economic Thought</i>	
J. R. Moroney: The Current State of Money and Production Theory.....	335	C. D. Goodwin: Economic Theory and Society: A Plea for Process Analysis.....	409
A. F. Brimmer: The Political Economy of Money: Evolution and Impact of Monetarism in the Federal Reserve System.....	344	H. G. Johnson: The Early Economics of Keynes.....	416
<i>Inequality: The Present Tendency and the Remedy</i>		<i>Economic Education</i>	
H. M. Hochman: Individual Preferences and Distributional Adjustments.....	353	A. C. Kelley: TIPS and Technical Change in Classroom Instruction.....	422
T. P. Schultz: Long-Term Changes in Personal Income Distribution: Theoretical Approaches, Evidence, and Explanations.....	361	R. Attiyeh and K. G. Lumsden: Some Modern Myths in Teaching Economics: The U.K. Experience.....	429
<i>The Economics of Full Racial Equality</i>		M. Zweig: Teaching Radical Political Economics in the Introductory Course.....	434
D. Bell: Occupational Discrimination as a Source of Income Differences: Lessons of the 1960's.....	363	<i>Long-Term Trends</i>	
<i>Economists Consider Economic Reporters and Vice Versa: A Discussion</i>		L. C. Thurow: The American Economy in the Year 2000.....	439
W. W. Heller: Introductory Comments.....	373	T. S. Khachaturov: Long-Term Planning and Forecasting in the USSR.....	444
		<i>Luncheon in Honor of Gunnar and Alva Myrdal</i>	
		G. Myrdal: Response to Introduction.....	456

## CONTRIBUTORS TO ARTICLES AND COMMUNICATIONS

- Alchian, A. A. 777  
 Allingham, M. G. 447  
 Arrow, K. J. 171  
 Bailey, D. 19  
 Baldwin, R. E. 465  
 Baron, D. P. 384  
 Baumol, W. J. 150, 175, 307  
 Behrman, J. R. 825  
 Bernhardt, I. 193  
 Bertrand, T. J. 743  
 Bhatia, K. B. 866  
 Bonomo, V. 000  
 Borcharding, T. E. 891  
 Borts, G. H. 964  
 Boskin, M. J. 466  
 Bradford, D. F. 175  
 Branson, W. H. 207  
 Breit, W. 501  
 Brittain, J. A. 739  
 Brubaker, E. R. 675, 997  
 Burmeister, E. 842  
 Chang, J. 221  
 Chetty, V. K. 226  
 Chiswick, B. R. 752  
 Cicchetti, C. J. 605  
 Coffman, R. B. 489  
 Cook, P. J. 139  
 Cukierman, A. 198  
 Culbertson, W. P. Jr. 501  
 Daly, G. 131  
 Darby, M. R. 928  
 Deacon, R. T. 891  
 Demsetz, H. 777  
 Diamond, P. A. 333  
 Douthett, C. M. 580  
 Dudley, L. 662  
 Eisner, R. 957  
 Fama, E. F. 670  
 Feige, E. L. 509  
 Feldstein, M. S. 32, 735  
 Ferber, M. A. 451  
 Fischer, S. 880  
 Fisher, A. C. 605  
 Flatters, F. 473  
 Galenson, M. 211  
 Gaumnitz, J. E. 438  
 Gayer, P. 714  
 Gibson, W. E. 854  
 Gierzt, F. 131  
 Goldberg, M. A. 692  
 Goldfarb, R. S. 714  
 Gordon, M. J. 107  
 Greenhut, M. L. 705  
 Griffin, J. M. 46  
 Grossman, H. I. 630  
 Hansen, W. L. 760  
 Harkness, J. 177  
 Harrison, B. 796  
 Hay, G. A. 403, 656  
 Hayami, Y. 119  
 Hines, A. G. 155  
 Holbrook, R. S. 57  
 Horst, T. 37  
 Jaffee, D. M. 484  
 Jay, K. 459  
 Juster, F. T. 564  
 Kain, J. F. 263  
 Kamien, M. I. 918  
 Katzner, D. W. 432  
 Kellman, M. 679  
 Kenen, P. B. 517  
 Keren, M. 149  
 Kessel, R. A. 527  
 Klevorick, A. K. 207  
 Kneese, A. V. 958  
 Koster, M. 323  
 Krutilla, J. V. 605  
 Kumar, P. 143  
 Laffer, A. B. 670  
 Laumas, P. S. 730  
 Lee, T. H. 217  
 Leland, H. E. 278  
 Lerner, A. P. 1009  
 Levhari, D. 185  
 Lind, R. C. 171  
 McCain, R. A. 497  
 McKean, R. N. 165  
 McManus, M. 489  
 Mankin, W. 980  
 Manove, M. 689  
 Masson, R. T. 77  
 Masters, S. 755  
 Mauer, L. J. 700  
 Means, G. C. 292  
 Melvin, J. R. 723  
 Mishan, E. J. 161, 971  
 Mohabbat, K. A. 730  
 Mohring, H. 591  
 Montias, J. M. 685  
 Mortensen, D. T. 718  
 Moore, J. H. 165  
 Mossin, J. 147  
 Nagatani, K. 344  
 Neher, P. 452, 979  
 Nelson, C. R. 902  
 Nichols, A. 168  
 Ng, Y-K. 173  
 Nordhaus, W. D. 428  
 Ohta, H. 705  
 Olson, J. E. 395  
 Paradis, G. E. 107  
 Parkin, J. M. 982  
 Patinkin, D. 185  
 Peltzman, S. 959  
 Perez, L. L. 679  
 Peterson, W. 119  
 Phelps, E. S. 533, 659  
 Phillips, L. 988

Philips, L. 747  
 Porter, R. B. 438  
 Quigley, J. M. 263  
 Ramanathan, R. 180  
 Reza, A. M. 1007  
 Ribich, T. 755  
 Robinson, W. C. 977  
 Rorke, C. H. 107  
 Rouzier, P. 747  
 Ruff, L. E. 186  
 Samuelson, P. A. 249  
 Sandmo, A. 194  
 Sawers, L. 414  
 Scadding, J. L. 151  
 Scanlon, W. J. 760  
 Scaperlanda, A. E. 700  
 Scherer, F. M. 422  
 Schmalensee, R. 813  
 Schotta, C. 19, 993  
 Schramm, R. 553  
 Schwartz, N. L. 918  
 Silberberg, E. 196, 942  
 Sims, C. A. 540  
 Smith, P. E. 230  
 Smith, V. L. 66, 477  
 Stein, J. L. 964  
 Steinhauer, L. 221  
 Stewart, D. B. 720  
 Stekler, H. O. 724

Stonebraker, R. J. 140  
 Tanner, J. E. 234  
 Taylor, L. D. 642  
 Tideman, T. N. 959  
 Tobin, J. 1  
 Travis, W. P. 87  
 Tsiang, S. C. 354  
 Tullock, G. 1004  
 Turnovsky, S. J. 842  
 Vandermeulen, D. C. 453  
 Veendorp, E. C. H. 1011  
 Vernon, J. R. 953  
 Wachtel, P. 564  
 Walton, G. M. 489  
 Waud, R. N. 1005  
 Waverman, L. 620  
 Weintraub, R. 988  
 Weisbrod, B. A. 760  
 Weiser, L. 459  
 Weiserbs, D. 642  
 Weiss, L. 372  
 Weitzman, M. L. 682  
 Welch, F. 323  
 Wellington, D. 170  
 Williamson, J. G. 372  
 Wu, S. Y. 982  
 Yaari, M. 333  
 Zajac, E. E. 142

## CONTRIBUTORS TO PAPERS AND PROCEEDINGS

Aaron, H. J. 332  
 Adams, W. 279  
 Adams, W. J. 279  
 Attiyeh, R. 429  
 Bell, D. 363  
 Berry, R. A. 403  
 Bradford, D. F. 99  
 Brimmer, A. F. 344  
 Browne, R. S. 39  
 Cleaver, Jr., H. M. 177  
 Currie, L. B. 139  
 Engle III, R. F. 87  
 Faine, H. R. 70  
 Fisher, F. M. 87  
 Fishlow, A. 391  
 Foster, M. I. 195  
 Friedman, M. 10  
 Friend, I. 212  
 Gintis, H. 267  
 Goodwin, C. D. 409  
 Gordon, D. M. 319  
 Gurley, J. G. 19  
 Haberler, G. 234  
 Harris, J. R. 87

Heller, W. H. 373  
 Hilton, G. W. 47  
 Hochman, H. M. 353  
 Janssen, R. M. 387  
 Johnson, H. G. 416  
 Jones, B. L. 125  
 Kaufman, R. 288  
 Kelley, A. C. 422  
 Keyserling, L. 134  
 Khachaturov, T. S. 444  
 Kurth, J. R. 304  
 Lampman, R. J. 331  
 Lekachman, R. 142  
 Lerner, A. P. 258  
 Lumsden, K. G. 429  
 Marris, R. 103  
 Marshall, F. R. 204  
 Melman, S. 312  
 Merewitz, L. 78  
 Meyer, J. R. 379  
 Modigliani, F. 229  
 Moroney, J. R. 335  
 Muth, R. F. 98  
 Myrdal, G. 456

Nathan, R. R. 138  
Nossiter, B. D. 381  
Okun, A. M. 24  
Phelps, C. D. 167  
Rasche, R. 220  
Reich, M. 296  
Robinson, J. 1  
Rothenberg, J. 87  
Sampson, R. J. 55  
Sandell, S. H. 175  
Schnittker, J. A. 389  
Schultz, T. P. 361  
Scitovsky, T. 62  
Sexton, B. 149

Silk, L. S. 376  
Strober, M. 154  
Sweezy, A. R. 116  
Tabb, W. K. 31  
Thompson, E. A. 329  
Thurrow, L. C. 439  
Ulman, L. 242  
Wachtel, H. 187  
Wallich, H. C. 384  
Weber, A. R. 251  
Weisskoff, F. B. 161  
Zellner, H. 157  
Zweig, M. 434





Number 73 of a series of photographs of past presidents of the Association



*James Tobin*

# Inflation and Unemployment

By JAMES TOBIN\*

The world economy today is vastly different from the 1930's, when Seymour Harris, the chairman of this meeting, infected me with his boundless enthusiasm for economics and his steadfast confidence in its capacity for good works. Economics is very different, too. Both the science and its subject have changed, and for the better, since World War II. But there are some notable constants. Unemployment and inflation still preoccupy and perplex economists, statesmen, journalists, housewives, and everyone else. The connection between them is the principal domestic economic burden of presidents and prime ministers, and the major area of controversy and ignorance in macroeconomics. I have chosen to review economic thought on this topic on this occasion, partly because of its inevitable timeliness, partly because of a personal interest reaching back to my first published work in 1941.

## I. The Meanings of Full Employment

Today, as thirty and forty years ago, economists debate how much unemployment is voluntary, how much involuntary; how much is a phenomenon of equilibrium, how much a symptom of disequilibrium; how much is compatible with competition, how much is to be blamed on monopolies, labor unions, and restrictive legislation; how much unemployment characterizes "full" employment.

Full employment—imagine macroeconomics deprived of the concept. But what is it? What is the proper employment goal of policies affecting aggregate de-

mand? Zero unemployment in the monthly labor force survey? That outcome is so inconceivable outside of Switzerland that it is useless as a guide to policy. Any other numerical candidate, yes even 4 percent, is patently arbitrary without reference to basic criteria. Unemployment equal to vacancies? Measurement problems aside, this definition has the same straightforward appeal as zero unemployment, which it simply corrects for friction.<sup>1</sup>

A concept of full employment more congenial to economic theory is labor market equilibrium, a volume of employment which is simultaneously the amount employers want to offer and the amount workers want to accept at prevailing wage rates and prices. Forty years ago theorists with confidence in markets could believe that full employment is whatever volume of employment the economy is moving toward, and that its achievement requires of the government nothing more than neutrality, and nothing less.

After Keynes challenged the classical notion of labor market equilibrium and the complacent view of policy to which it led, full employment came to mean maximum aggregate supply, the point at which expansion of aggregate demand could not further increase employment and output.

Full employment was also regarded as the economy's inflation threshold. With a deflationary gap, demand less than full employment supply, prices would be declining or at worst constant. Expansion of aggregate demand short of full employment would cause at most a one-shot

\* Presidential address delivered at the eighty-fourth meeting of the American Economic Association, New Orleans, Louisiana, December 28, 1971.

<sup>1</sup> This concept is commonly attributed to W. H. Beveridge, but he was actually more ambitious and required a surplus of vacancies.

increase of prices. For continuing inflation, the textbooks told us, a necessary and sufficient condition was an inflationary gap, real aggregate demand in excess of feasible supply. The model was tailor-made for wartime inflation.

Postwar experience destroyed the identification of full employment with the economy's inflation threshold. The profession, the press, and the public discovered the "new inflation" of the 1950's, inflation without benefit of gap, labelled but scarcely illuminated by the term "cost-push." Subsequently the view of the world suggested by the Phillips curve merged demand-pull and cost-push inflation and blurred the distinction between them. This view contained no concept of full employment. In its place came the tradeoff, along which society supposedly can choose the least undesirable feasible combination of the evils of unemployment and inflation.

Many economists deny the existence of a durable Phillips tradeoff. Their numbers and influence are increasing. Some of them contend that there is only one rate of unemployment compatible with steady inflation, a "natural rate" consistent with any steady rate of change of prices, positive, zero, or negative. The natural rate is another full employment candidate, a policy target at least in the passive sense that monetary and fiscal policy makers are advised to eschew any numerical unemployment goal and to let the economy gravitate to this equilibrium. So we have come full circle. Full employment is once again nothing but the equilibrium reached by labor markets unaided and undistorted by governmental fine tuning.

In discussing these issues, I shall make the following points. First, an observed amount of unemployment is not revealed to be voluntary simply by the fact that money wage rates are constant, or rising, or even accelerating. I shall recall and extend Keynes's definition of involuntary

unemployment and his explanation why workers may accept price inflation as a method of reducing real wages while rejecting money wage cuts. The second point is related. Involuntary unemployment is a disequilibrium phenomenon; the behavior, the persistence, of excess supplies of labor depend on how and how fast markets adjust to shocks, and on how large and how frequent the shocks are. Higher prices or faster inflation can diminish involuntary, disequilibrium unemployment, even though voluntary, equilibrium labor supply is entirely free of money illusion.

Third, various criteria of full employment coincide in a theoretical full stationary equilibrium, but diverge in persistent disequilibrium. These are 1) the natural rate of unemployment, the rate compatible with zero or some other constant inflation rate, 2) zero involuntary unemployment, 3) the rate of unemployment needed for optimal job search and placement, and 4) unemployment equal to job vacancies. The first criterion dictates higher unemployment than any of the rest. Instead of commending the natural rate as a target of employment policy, the other three criteria suggest less unemployment and more inflation. Therefore, fourth, there are real gains from additional employment, which must be weighed in the social balance against the costs of inflation. I shall conclude with a few remarks on this choice, and on the possibilities of improving the terms of the tradeoff.

## II. Keynesian and Classical Interpretations of Unemployment

To begin with the *General Theory* is not just the ritual piety economists of my generation owe the book that shaped their minds. Keynes's treatment of labor market equilibrium and disequilibrium in his first chapter is remarkably relevant today.

Keynes attacked what he called the classical presumption that persistent unemployment is voluntary unemployment. The presumption he challenged is that in competitive labor markets actual employment and unemployment reveal workers' true preferences between work and alternative uses of time, the presumption that no one is fully or partially unemployed whose real wage per hour exceeds his marginal valuation of an hour of free time. Orthodox economists found the observed stickiness of money wages to be persuasive evidence that unemployment, even in the Great Depression, was voluntary. Keynes found decisive evidence against this inference in the willingness of workers to accept a larger volume of employment at a lower real wage resulting from an increase of prices.

Whenever unemployment could be reduced by expansion of aggregate demand, Keynes regarded it as involuntary. He expected expansion to raise prices and lower real wages, but this expectation is not crucial to his argument. Indeed, if it is possible to raise employment without reduction in the real wage, his case for calling the unemployment involuntary is strengthened.

But why is the money wage so stubborn if more labor is willingly available at the same or lower real wage? Consider first some answers Keynes did not give. He did not appeal to trade union monopolies or minimum wage laws. He was anxious, perhaps over-anxious, to meet his putative classical opponents on their home field, the competitive economy. He did not rely on any failure of workers to perceive what a rise in prices does to real wages. The unemployed take new jobs, the employed hold old ones, with eyes open. Otherwise the new situation would be transient.

Instead, Keynes emphasized the institutional fact that wages are bargained and set in the monetary unit of account. Money wage rates are, to use an unKeynes-

ian term, "administered prices." That is, they are not set and reset in daily auctions but posted and fixed for finite periods of time. This observation led Keynes to his central explanation: Workers, individually and in groups, are more concerned with relative than absolute real wages. They may withdraw labor if their wages fall relatively to wages elsewhere, even though they would not withdraw any if real wages fall uniformly everywhere. Labor markets are decentralized, and there is no way money wages can fall in any one market without impairing the relative status of the workers there. A general rise in prices is a neutral and universal method of reducing real wages, the only method in a decentralized and uncontrolled economy. Inflation would not be needed, we may infer, if by government compulsion, economy-wide bargaining, or social compact, all money wage rates could be scaled down together.

Keynes apparently meant that relative wages are the arguments in labor supply functions. But Alchian (pp. 27-52 in Phelps et al.) and other theorists of search activity have offered a somewhat different interpretation, namely that workers whose money wages are reduced will quit their jobs to seek employment in other markets where they think, perhaps mistakenly, that wages remain high.

Keynes's explanation of money wage stickiness is plausible and realistic. But two related analytical issues have obscured the message. Can there be involuntary unemployment in an equilibrium, a proper, full-fledged neoclassical equilibrium? Does the labor supply behavior described by Keynes betray "money illusion"? Keynes gave a loud yes in answer to the first question, and this seems at first glance to compel an affirmative answer to the second.

An economic theorist can, of course, commit no greater crime than to assume money illusion. Comparative statics is a

nonhistorical exercise, in which different price levels are to be viewed as alternative rather than sequential. Compare two situations that differ only in the scale of exogenous monetary variables; imagine, for example, that all such magnitudes are ten times as high in one situation as in the other. All equilibrium prices, including money wage rates, should differ in the same proportion, while all real magnitudes, including employment, should be the same in the two equilibria. To assume instead that workers' supply decisions vary with the price level is to say that they would behave differently if the unit of account were, and always had been, dimes instead of dollars. Surely Keynes should not be interpreted to attribute to anyone money illusion in this sense. He was not talking about so strict and static an equilibrium.

Axel Leijonhufvud's illuminating and perceptive interpretation of Keynes argues convincingly that, in chapter 1 as throughout the *General Theory*, what Keynes calls equilibrium should be viewed as persistent disequilibrium, and what appears to be comparative statics is really shrewd and incisive, if awkward, dynamic analysis. Involuntary unemployment means that labor markets are not in equilibrium. The resistance of money wage rates to excess supply is a feature of the adjustment process rather than a symptom of irrationality.

The other side of Keynes's story is that in depressions money wage deflation, even if it occurred more speedily, or especially if it occurred more speedily, would be at best a weak equilibrator and quite possibly a source of more unemployment rather than less. In contemporary language, the perverse case would arise if a high and ever-increasing real rate of return on money inhibited real demand faster than the rising purchasing power of monetary stocks stimulated demand. To pursue this Keynesian theme further here would be a digression.

What relevance has this excursion into depression economics for contemporary problems of unemployment and wage inflation? The issues are remarkably similar, even though events and Phillips have shifted attention from levels to time rates of change of wages and prices. Phillips curve doctrine<sup>2</sup> is in an important sense the postwar analogue of Keynesian wage and employment theory, while natural rate doctrine is the contemporary version of the classical position Keynes was opposing.

Phillips curve doctrine implies that lower unemployment can be purchased at the cost of faster inflation. Let us adapt Keynes's test for involuntary unemployment to the dynamic terms of contemporary discussion of inflation, wages, and unemployment. Suppose that the current rate of unemployment continues. Associated with it is a path of real wages, rising at the rate of productivity growth. Consider an alternative future, with unemployment at first declining to a rate one percentage point lower and then remaining constant at the lower rate. Associated with the lower unemployment alternative will be a second path of real wages. Eventually this real wage path will show, at least to first approximation, the same rate of increase as the first one, the rate of productivity growth. But the paths may differ because of the transitional effects of increasing the rate of employment. The growth of real wages will be retarded in the short run if additional employment lowers labor's marginal productivity. In any case, the test question is whether with full information about the two alternatives labor would accept the second one—

<sup>2</sup> Phillips himself is not a prophet of the doctrine associated with his curve. His 1958 article was probably the most influential macro-economic paper of the last quarter century. But Phillips simply presented some striking empirical findings, which others have replicated many times for many economies. He is not responsible for the theories and policy conclusions his findings

whether, in other words, the additional employment would be willingly supplied along the second real wage path. If the answer is affirmative, then that one percentage point of unemployment is involuntary.

For Keynes's reasons, a negative answer cannot necessarily be inferred from failure of money wage rates to fall or even decelerate. Actual unemployment and the real wage path associated with it are not necessarily an equilibrium. Rigidities in the path of money wage rates can be explained by workers' preoccupation with relative wages and the absence of any central economy-wide mechanism for altering all money wages together.

According to the natural rate hypothesis, there is just one rate of unemployment compatible with steady wage and price inflation, and this is in the long run compatible with any constant rate of change of prices, positive, zero, or negative. Only at the natural rate of unemployment are workers content with current and prospective real wages, content to have their real wages rise at the rate of growth of productivity. Along the feasible path of real wages they would not wish to accept any larger volume of employment. Lower unemployment, therefore, can arise only from economy-wide excess demand for labor and must generate a gap between real wages desired and real wages earned. The gap evokes increases of money wages designed to raise real wages faster than productivity. But this intention is always frustrated, the gap is never closed, money wages and prices accelerate. By symmetrical argument, unemployment above the natural rate signifies excess supply in labor markets and ever accelerating deflation. Older classical economists regarded constancy of money wage rates as indicative of full employment equilibrium, at which the allocation of time between work and other pursuits is revealed as voluntary

same claims for the natural rate of unemployment, except that in the equilibrium money wages are not necessarily constant but growing at the rate of productivity gain plus the experienced and expected rate of inflation of prices.

### III. Is Zero-Inflation Unemployment Voluntary and Optimal?

There are, then, two conflicting interpretations of the welfare value of employment in excess of the level consistent with price stability. One is that additional employment does not produce enough to compensate workers for the value of other uses of their time. The fact that it generates inflation is taken as *prima facie* evidence of a welfare loss. The alternative view, which I shall argue, is that the responses of money wages and prices to changes in aggregate demand reflect mechanics of adjustment, institutional constraints, and relative wage patterns and reveal nothing in particular about individual or social valuations of unemployed time vis-à-vis the wages of employment.

On this rostrum four years ago, Milton Friedman identified the noninflationary natural rate of unemployment with "equilibrium in the structure of real wage rates" (p. 8). "The 'natural rate of unemployment,'" he said, "... is the level that would be ground out by the Walrasian system of general equilibrium equations, provided that there is embedded in them the actual structural characteristics of the labor and commodity markets, including market imperfections, stochastic variability in demands and supplies, the costs of getting information about job vacancies and labor availabilities, the costs of mobility, and so on." Presumably this Walrasian equilibrium also has the usual optimal properties; at any rate, Friedman advised the monetary authorities not to seek to improve upon it. But in fact we

Walrasian equilibrium that allows for all the imperfections and frictions that explain why the natural rate is bigger than zero, and even less about the optimality of such an equilibrium if it exists.

In the new microeconomics of labor markets and inflation, the principal activity whose marginal value sets the reservation price of employment is job search. It is not pure leisure, for in principle persons who choose that option are not reported as unemployed; however, there may be a leisure component in job seeking.

A crucial assumption of the theory is that search is significantly more efficient when the searcher is unemployed, but almost no evidence has been advanced on this point. Members of our own profession are adept at seeking and finding new jobs without first leaving their old ones or abandoning not-in-labor-force status. We do not know how many quits and new hires in manufacturing are similar transfers, but some of them must be; if all reported accessions were hires of unemployed workers, the mean duration of unemployment would be only about half what it is in fact. In surveys of job mobility among blue collar workers in 1946-47 (see Lloyd Reynolds, pp. 214-15, and Herbert Parnes, pp. 158-59), 25 percent of workers who quit had new jobs lined up in advance. Reynolds found that the main obstacle to mobility without unemployment was not lack of information or time, but simply "anti-pirating" collusion by employers.

A considerable amount of search activity by unemployed workers appears to be an unproductive consequence of dissatisfaction and frustration rather than a rational quest for improvement. This was the conclusion of Reynolds' survey twenty-five years ago, p. 215, and it has been re-emphasized for the contemporary scene by Robert Hall, and by Peter Doeringer and Michael Piore for what they term the secondary labor force. Reynolds found

that quitting a job to look for a new one while unemployed actually yielded a better job in only a third of the cases. Lining up a new job in advance was a more successful strategy: two-thirds of such changes turned out to be improvements. Today, according to the dual labor market hypothesis, the basic reason for frequent and long spells of unemployment in the secondary labor force is the shortage of good jobs.

In any event, the contention of some natural rate theorists is that employment beyond the natural rate takes time that would be better spent in search activity. Why do workers accept such employment? An answer to this question is a key element in a theory that generally presumes that actual behavior reveals true preferences. The answer given is that workers accept the additional employment only because they are victims of inflation illusion. One form of inflation illusion is over-estimation of the real wages of jobs they now hold, if they are employed, or of jobs they find, if they are unemployed and searching. If they did not under-estimate price inflation, employed workers would more often quit to search, and unemployed workers would search longer.

The force of this argument seems to me diluted by the fact that price inflation illusion affects equally both sides of the job seeker's equation. He over-estimates the real value of an immediate job, but he also over-estimates the real values of jobs he might wait for. It is in the spirit of this theorizing to assume that money interest rates respond to the same correct or incorrect inflationary expectations. As a first approximation, inflation illusion has no substitution effect on the margin between working and waiting.

It does have an income effect, causing workers to exaggerate their real wealth. In which direction the income effect would work is not transparent. Does greater wealth, or the illusion of greater

wealth, make people more choosy about jobs, more inclined to quit and to wait? Or less choosy, more inclined to stay in the job they have or to take the first one that comes along? I should have thought more selective rather than less. But natural rate theory must take the opposite view if it is to explain why under-estimation of price inflation bamboozles workers into holding or taking jobs that they do not really want.

Another form of alleged inflation illusion refers to wages rather than prices. Workers are myopic and do not perceive that wages elsewhere are, or soon will be, rising as fast as the money wage of the job they now hold or have just found. Consequently they under-estimate the advantages of quitting and searching. This explanation is convincing only to the extent that the payoff to search activity is determined by wage differentials. The payoff also depends on the probabilities of getting jobs at quoted wages, therefore on the balance between vacancies and job seekers. Workers know that perfectly well. Quit rates are an index of voluntary search activity. They do not diminish when unemployment is low and wage rates are rapidly rising. They increase, quite understandably. This fact contradicts the inflation illusion story, both versions. I conclude that it is not possible to regard fluctuations of unemployment on either side of the zero-inflation rate as mainly voluntary, albeit mistaken, extensions and contractions of search activity.

The new microeconomics of job search (see Edmund Phelps et al.), is nevertheless a valuable contribution to understanding of frictional unemployment. It provides reasons why some unemployment is voluntary, and why some unemployment is socially efficient.

Does the market produce the *optimal* amount of search unemployment? Is the natural rate optimal? I do not believe the

new microeconomics has yet answered these questions.

An omniscient and beneficent economic dictator would not place every new job seeker immediately in any job at hand. Such a policy would create many mismatches, sacrificing efficiency in production or necessitating costly job-to-job shifts later on. The hypothetical planner would prefer to keep a queue of workers unemployed, so that he would have a larger choice of jobs to which to assign them. But he would not make the queue too long, because workers in the queue are not producing anything.

Of course he could shorten the queue of unemployed if he could dispose of more jobs and lengthen the queue of vacancies. With enough jobs of various kinds, he would never lack a vacancy for which any worker who happens to come along has comparative advantage. But because of limited capital stocks and interdependence among skills, jobs cannot be indefinitely multiplied without lowering their marginal productivity. Our wise and benevolent planner would not place people in jobs yielding less than the marginal value of leisure. Given this constraint on the number of jobs, he would always have to keep some workers waiting, and some jobs vacant. But he certainly would be inefficient if he had fewer jobs, filled and vacant, than this constraint. This is the common sense of Beveridge's rule—that vacancies should not be less than unemployment.

Is the natural rate a market solution of the hypothetical planner's operations research problem? According to search theory, an unemployed worker considers the probabilities that he can get a better job by searching longer and balances the expected discounted value of waiting against the loss of earnings. The employed worker makes a similar calculation when he considers quitting, also taking into ac-

count the once and for all costs of movement. These calculations are like those of the planner, but with an important difference. An individual does not internalize all the considerations the planner takes into account. The external effects are the familiar ones of congestion theory. A worker deciding to join a queue or to stay in one considers the probabilities of getting a job, but not the effects of his decision on the probabilities that others face. He lowers those probabilities for people in the queue he joins and raises them for persons waiting for the kind of job he vacates or turns down. Too many persons are unemployed waiting for good jobs, while less desirable ones go begging. However, external effects also occur in the decisions of employers whether to fill a vacancy with the applicant at hand or to wait for someone more qualified. It is not obvious, at least to me, whether the market is biased toward excessive or inadequate search. But it is doubtful that it produces the optimal amount.

Empirically the proposition that in the United States the zero-inflation rate of unemployment reflects voluntary and efficient job-seeking activity strains credulity. If there were a natural rate of unemployment in the United States, what would it be? It is hard to say because virtually all econometric Phillips curves allow for a whole menu of steady inflation rates. But estimates constrained to produce a vertical long-run Phillips curve suggest a natural rate between 5 and 6 percent of the labor force.<sup>3</sup>

So let us consider some of the features of an overall unemployment rate of 5 to 6 percent. First, about 40 percent of accessions in manufacturing are rehires rather than new hires. Temporarily laid off by their employers, these workers had been awaiting recall and were scarcely engaged in

voluntary search activity. Their unemployment is as much a deadweight loss as the disguised unemployment of redundant workers on payrolls. This number declines to 25–30 percent when unemployment is 4 percent or below. Likewise, a 5–6 percent unemployment rate means that voluntary quits amount only to about a third of separations, layoffs to two-thirds. The proportions are reversed at low unemployment rates.

Second, the unemployment statistic is not an exhaustive count of those with time and incentive to search. An additional 3 percent of the labor force are involuntarily confined to part-time work, and another  $3/4$  of 1 percent are out of the labor force because they “could not find job” or “think no work available”—discouraged by market conditions rather than personal incapacities.

Third, with unemployment of 5–6 percent the number of reported vacancies is less than  $1/2$  of 1 percent. Vacancies appear to be understated relative to unemployment, but they rise to  $1\frac{1}{2}$  percent when the unemployment rate is below 4 percent. At 5–6 percent unemployment, the economy is clearly capable of generating many more jobs with marginal productivity high enough so that people prefer them to leisure. The capital stock is no limitation, since 5–6 percent unemployment has been associated with more than 20 percent excess capacity. Moreover, when more jobs are created by expansion of demand, with or without inflation, labor force participation increases; this would hardly occur if the additional jobs were low in quality and productivity. As the parable of the central employment planner indicates, there will be excessive waiting for jobs if the roster of jobs and the menu of vacancies are suboptimal.

In summary, labor markets characterized by 5–6 percent unemployment do not display the symptoms one would ex-

<sup>3</sup> See Lucas and Rapping, pp. 257–305, in Phelps et al.

pect if the unemployment were voluntary search activity. Even if it were voluntary, search activity on such a large scale would surely be socially wasteful. The only reason anyone might regard so high an unemployment rate as an equilibrium and social optimum is that lower rates cause accelerating inflation. But this is almost tautological. The inferences of equilibrium and optimality would be more convincing if they were corroborated by direct evidence.

#### IV. Why is There Inflation without Aggregate Excess Demand?

Zero-inflation unemployment is not wholly voluntary, not optimal, I might even say not natural. In other words, the economy has an inflationary bias: When labor markets provide as many jobs as there are willing workers, there is inflation, perhaps accelerating inflation. Why?

The Phillips curve has been an empirical finding in search of a theory, like Pirandello characters in search of an author. One rationalization might be termed a theory of stochastic macro-equilibrium: stochastic, because random intersectoral shocks keep individual labor markets in diverse states of disequilibrium; macro-equilibrium, because the perpetual flux of particular markets produces fairly definite aggregate outcomes of unemployment and wages. Stimulated by Phillips's 1958 findings, Richard Lipsey proposed a model of this kind in 1960, and it has since been elaborated by Archibald, pp. 212-23 and Holt, pp. 53-123 and 224-56 in Phelps et. al., and others. I propose now to sketch a theory in the same spirit.

It is an essential feature of the theory that economy-wide relations among employment, wages, and prices are aggregations of diverse outcomes in heterogeneous markets. The myth of macroeconomics is that relations among aggregates are en-

larged analogues of relations among corresponding variables for individual households, firms, industries, markets. The myth is a harmless and useful simplification in many contexts, but sometimes it misses the essence of the phenomenon.

Unemployment is, in this model as in Keynes reinterpreted, a disequilibrium phenomenon. Money wages do not adjust rapidly enough to clear all labor markets every day. Excess supplies in labor markets take the form of unemployment, and excess demands the form of unfilled vacancies. At any moment, markets vary widely in excess demand or supply, and the economy as a whole shows both vacancies and unemployment.

The overall balance of vacancies and unemployment is determined by aggregate demand, and is therefore in principle subject to control by overall monetary and fiscal policy. Higher aggregate demand means fewer excess supply markets and more excess demand markets, accordingly less unemployment and more vacancies.

In any particular labor market, the rate of increase of money wages is the sum of two components, an equilibrium component and a disequilibrium component. The first is the rate at which the wage would increase were the market in equilibrium, with neither vacancies nor unemployment. The other component is a function of excess demand and supply—a monotonic function, positive for positive excess demand, zero for zero excess demand, non-positive for excess supply. I begin with the disequilibrium component.

Of course the disequilibrium components are relevant only if disequilibria persist. Why aren't they eliminated by the very adjustments they set in motion? Workers will move from excess supply markets to excess demand markets, and from low wage to high wage markets. Unless they overshoot, these movements are equilibrating. The theory therefore

requires that new disequilibria are always arising. Aggregate demand may be stable, but beneath its stability is never-ending flux: new products, new processes, new tastes and fashions, new developments of land and natural resources, obsolescent industries and declining areas.

The overlap of vacancies and unemployment—say, the sum of the two for any given difference between them—is a measure of the heterogeneity or dispersion of individual markets. The amount of dispersion depends directly on the size of those shocks of demand and technology that keep markets in perpetual disequilibrium, and inversely on the responsive mobility of labor. The one increases, the other diminishes the frictional component of unemployment, that is, the number of unfilled vacancies coexisting with any given unemployment rate.

A central assumption of the theory is that the functions relating wage change to excess demand or supply are non-linear, specifically that unemployment retards money wages less than vacancies accelerate them. Nonlinearity in the response of wages to excess demand has several important implications. First, it helps to explain the characteristic observed curvature of the Phillips curve. Each successive increment of unemployment has less effect in reducing the rate of inflation. Linear wage response, on the other hand, would mean a linear Phillips relation.

Second, given the overall state of aggregate demand, economy-wide vacancies less unemployment, wage inflation will be greater the larger the variance among markets in excess demand and supply. As a number of recent empirical studies, have confirmed (see George Perry and Charles Schultze), dispersion is inflationary. Of course, the rate of wage inflation will depend not only on the overall dispersion of excess demands and supplies across markets but also on the

particular markets where the excess supplies and demands happen to fall. An unlucky random drawing might put the excess demands in highly responsive markets and the excess supplies in especially unresponsive ones.

Third, the nonlinearity is an explanation of inflationary bias, in the following sense. Even when aggregate vacancies are at most equal to unemployment, the average disequilibrium component will be positive. Full employment in the sense of equality of vacancies and unemployment is not compatible with price stability. Zero inflation requires unemployment in excess of vacancies.

Criteria that coincide in full long-run equilibrium—zero inflation and zero aggregate excess demand—diverge in stochastic macro-equilibrium. Full long-run equilibrium in all markets would show no unemployment, no vacancies, no unanticipated inflation. But with unending sectoral flux, zero excess demand spells inflation and zero inflation spells net excess supply, unemployment in excess of vacancies. In these circumstances neither criterion can be justified simply because it is a property of full long-run equilibrium. Both criteria automatically allow for frictional unemployment incident to the required movements of workers between markets; the no-inflation criterion requires enough additional unemployment to wipe out inflationary bias.

I turn now to the equilibrium component, the rate of wage increase in a market with neither excess demand nor excess supply. It is reasonable to suppose that the equilibrium component depends on the trend of wages of comparable labor elsewhere. A “competitive wage,” one that reflects relevant trends fully, is what employers will offer if they wish to maintain their share of the volume of employment. This will happen where the rate of growth of marginal revenue product—the com-

pound of productivity increase and price inflation—is the same as the trend in wages. But in some markets the equilibrium wage will be rising faster, and in others slower, than the economy-wide wage trend.

A “natural rate” result follows if actual wage increases feed fully into the equilibrium components of future wage increases. There will be acceleration whenever the non-linear disequilibrium effects are on average positive, and steady inflation, that is stochastically steady inflation, only at unemployment rates high enough to make the disequilibrium effects wash out. Phillips tradeoffs exist in the short run, and the time it takes for them to evaporate depends on the lengths of the lags with which today’s actual wage gains become tomorrow’s standards.

A rather minor modification may preserve Phillips tradeoffs in the long run. Suppose there is a floor on wage change in excess supply markets, independent of the amount of excess supply and of the past history of wages and prices. Suppose, for example, that wage change is never negative; it is either zero or what the response function says, whichever is algebraically larger. So long as there are markets where this floor is effective, there can be determinate rates of economy-wide wage inflation for various levels of aggregate demand. Markets at the floor do not increase their contributions to aggregate wage inflation when overall demand is raised. Nor is their contribution escalated to actual wage experience. But the frequency of such markets diminishes, it is true, both with overall demand and with inflation. The floor phenomenon can preserve a Phillips tradeoff within limits, but one that becomes ever more fragile and vanishes as greater demand pressure removes markets from contact with the zero floor. The model implies a long-run Phillips curve that is very flat for high unemployment

and becomes vertical at a critically low rate of unemployment.

These implications seem plausible and even realistic. It will be objected, however, that any permanent floor independent of general wage and price history and expectation must indicate money illusion. The answer is that the floor need not be permanent in any single market. It could give way to wage reduction when enough unemployment has persisted long enough. But with stochastic intersectoral shifts of demand, markets are always exchanging roles, and there can always be some markets, not always the same ones, at the floor.

This model avoids the empirically questionable implication of the usual natural rate hypothesis that unemployment rates only slightly higher than the critical rate will trigger ever-accelerating deflation. Phillips curves seem to be pretty flat at high rates of unemployment. During the great contraction of 1930–33, wage rates were slow to give way even in the face of massive unemployment and substantial deflation in consumer prices. Finally in 1932 and 1933 money wage rates fell more sharply, in response to prolonged unemployment, layoffs, shutdowns, and to threats and fears of more of the same.

I have gone through this example to make the point that irrationality, in the sense that meaningless differences in money values *permanently* affect individual behavior, is not logically necessary for the existence of a long-run Phillips tradeoff. In full long-run equilibrium in all markets, employment and unemployment would be independent of the levels and rates of change of money wage rates and prices. But this is not an equilibrium that the system ever approaches. The economy is in perpetual sectoral disequilibrium even when it has settled into a stochastic macro-equilibrium.

I suppose that one might maintain that asymmetry in wage adjustment and tem-

porary resistance to money wage decline reflect money illusion in some sense. Such an assertion would have to be based on an extension of the domain of well-defined rational behavior to cover responses to change, adjustment speeds, costs of information, costs of organizing and operating markets, and a host of other problems in dynamic theory. These theoretical extensions are in their infancy, although much work of interest and promise is being done. Meanwhile, I doubt that significant restrictions on disequilibrium adjustment mechanisms can be deduced from first principles.

Why are the wage and salary rates of employed workers so insensitive to the availability of potential replacements? One reason is that the employer makes some explicit or implicit commitments in putting a worker on the payroll in the first place. The employee expects that his wages and terms of employment will steadily improve, certainly never retrogress. He expects that the employer will pay him the rate prevailing for persons of comparable skill, occupation, experience, and seniority. He expects such commitments in return for his own investments in the job; arrangements for residence, transportation, and personal life involve set-up costs which will be wasted if the job turns sour. The market for labor services is not like a market for fresh produce where the entire current supply is auctioned daily. It is more like a rental housing market, in which most existing tenancies are the continuations of long-term relationships governed by contracts or less formal understandings.

Employers and workers alike regard the wages of comparable labor elsewhere as a standard, but what determines those reference wages? There is not even an auction where workers and employers unbound by existing relationships and commitments meet and determine a market-clearing wage. If such markets existed, they would

provide competitively determined guides for negotiated and administered wages, just as stock exchange prices are reference points for stock transactions elsewhere. In labor markets the reverse is closer to the truth. Wage rates for existing employees set the standards for new employees, too.

The equilibrium components of wage increases, it has been argued, depend on past wage increases throughout the economy. In those theoretical and econometric models of inflation where labor markets are aggregated into a single market, this relationship is expressed as an autoregressive equation of fixed structure: current wage increase depends on past wage increases. The same description applies when past wage increases enter indirectly, mediated by price inflation and productivity change. The process of mutual interdependence of market wages is a good deal more complex and less mechanical than these aggregated models suggest.

Reference standards for wages differ from market to market. The equilibrium wage increase in each market will be some function of past wages in all markets, and perhaps of past prices too. But the function need not be the same in every market. Wages of workers contiguous in geography, industry, and skill will be heavily weighted. Imagine a wage pattern matrix of coefficients describing the dependence of the percentage equilibrium wage increase in each market on the past increases in all other markets. The coefficients in each row are non-negative and sum to one, but their distribution across markets and time lags will differ from row to row.

Consider the properties of such a system in the absence of disequilibrium inputs. First, the system has the "natural rate" property that its steady state is indeterminate. Any rate of wage increase that has been occurring in all markets for a long enough time will continue. Second, from irregular initial conditions the system will

move toward one of these steady states, but which one depends on the specifics of the wage pattern matrix and the initial conditions. Contrary to some pessimistic warnings, there is no arithmetic compulsion that makes the whole system gravitate in the direction of its most inflationary sectors. The ultimate steady state inflation will be at most that of the market with the highest initial inflation rate, and at least that of the market with the lowest initial inflation rate. It need not be equal to the average inflation rate at the beginning, but may be either greater or smaller. Third, the adjustment paths are likely to contain cyclical components, damped or at most of constant amplitude, and during adjustments both individual and average wage movements may diverge substantially in both directions from their ultimate steady state value. Fourth, since wage decisions and negotiations occur infrequently, relative wage adjustments involve a lot of catching up and leap-frogging, and probably take a long time. I have sketched the formal properties of a disaggregated wage pattern system of this kind simply to stress again the vast simplification of the one-market myth.

A system in which only relative magnitudes matter has only a neutral equilibrium, from which it can be permanently displaced by random shocks. Even when a market is in equilibrium, it may outdo the recent wage increases in related markets. A shock of this kind, even though it is not repeated, raises permanently the steady state inflation rate. This is true cost-push—inflation generated neither by previous inflation nor by current excess demand. Shocks, of course, may be negative as well as positive. For example, upward pushes arising from adjustments in relative wage *levels* will be reversed when those adjustments are completed.

To the extent that one man's reference wages are another man's wages, there is

something arbitrary and conventional, indeterminate and unstable, in the process of wage setting. In the same current market circumstances, the reference pattern might be 8 percent per year or 3 percent per year or zero, depending on the historical prelude. Market conditions, unemployment and vacancies and their distributions, shape history and alter reference patterns. But accidental circumstances affecting strategic wage settlements also cast a long shadow.

Price inflation, as previously observed, is a neutral method of making arbitrary money wage paths conform to the realities of productivity growth, neutral in preserving the structure of relative wages. If expansion of aggregate demand brings both more inflation and more employment, there need be no mystery why unemployed workers accept the new jobs, or why employed workers do not vacate theirs. They need not be victims of ignorance or inflation illusion. They genuinely want more work at feasible real wages, and they also want to maintain the relative status they regard as proper and just.

Guideposts could be in principle the functional equivalent of inflation, a neutral method of reconciling wage and productivity paths. The trick is to find a formula for mutual deescalation which does not offend conceptions of relative equity. No one has devised a way of controlling average wage rates without intervening in the competitive struggle over relative wages. Inflation lets this struggle proceed and blindly, impartially, impersonally, and nonpolitically scales down all its outcomes. There are worse methods of resolving group rivalries and social conflict.

## V. The Role of Monopoly Power

Probably the most popular explanation of the inflationary bias of the economy is concentration of economic power in large corporations and unions. These powerful

monopolies and oligopolies, it is argued, are immune from competition in setting wages and prices. The unions raise wages above competitive rates, with little regard for the unemployed and under-employed workers knocking at the gates. Perhaps the unions are seeking a bigger share of the revenues of the monopolies and oligopolies with whom they bargain. But they don't really succeed in that objective, because the corporations simply pass the increased labor costs, along with mark-ups, on to their helpless customers. The remedy, it is argued, is either atomization of big business and big labor or strict public control of their prices and wages.

So simple a diagnosis is vitiated by confusion between levels and rates of change. Monopoly power is no doubt responsible for the relatively high prices and wages of some sectors. But can the exercise of monopoly power generate ever-rising price and wages? Monopolists have no reason to hold reserves of unexploited power. But if they did, or if events awarded them new power, their exploitation of it would raise their real prices and wages only temporarily.

Particular episodes of inflation may be associated with accretions of monopoly power, or with changes in the strategies and preferences of those who possess it. Among the reasons that wages and prices rose in the face of mass unemployment after 1933 were *NRA* codes and other early New Deal measures to suppress competition, and the growth of trade union membership and power under the protection of new federal legislation. Recently we have witnessed substantial gains in the powers of organized public employees. Unions elsewhere may not have gained power, but some of them apparently have changed their objectives in favor of wages at the expense of employment.

One reason for the popularity of the monopoly power diagnosis of inflation is

the identification of administered prices and wages with concentrations of economic power. When price and wage increases are the outcomes of visible negotiations and decisions, it seems obvious that identifiable firms and unions have the power to affect the course of inflation. But the fact that monopolies, oligopolies, and large unions have discretion does not mean it is invariably to their advantage to use it to raise prices and wages. Nor are administered prices and wages found only in high concentration sectors. Very few prices and wages in a modern economy, even in the more competitive sectors, are determined in Walrasian auction markets.

No doubt there has been a secular increase in the prevalence of administered wages and prices, connected with the relative decline of agriculture and other sectors of self-employment. This development probably has contributed to the inflationary bias of the economy, by enlarging the number of labor markets where the response of money wages to excess supply is slower than their response to excess demand. The decline of agriculture as a sector of flexible prices and wages and as an elastic source of industrial labor is probably an important reason why the Phillips trade off problem is worse now than in the 1920's. Sluggishness of response to excess supply is a feature of administered prices, whatever the market structure, but it may be accentuated by concentration of power per se. For example, powerful unions, not actually forced by competition to moderate their wage demands, may for reasons of internal politics be slow to respond to unemployment in their ranks.

## VI. Some Reflections on Policy

If the makers of macro-economic policy could be sure that the zero-inflation rate of unemployment is natural, voluntary, and optimal, their lives would be easy.

Friedman told us that all macro-economic policy needs to do, all it should try to do, is to make nominal national income grow steadily at the natural rate of growth of aggregate supply. This would sooner or later result in price stability. Steady price deflation would be even better, he said, because it would eliminate the socially wasteful incentive to economize money holdings. In either case, unemployment will converge to its natural rate, and wages and prices will settle into steady trends. Under this policy, whatever unemployment the market produces is the correct result. No tradeoff, no choice, no agonizing decisions.

I have argued this evening that a substantial amount of the unemployment compatible with zero inflation is involuntary and nonoptimal. This is, in my opinion, true whether or not the inflations associated with lower rates of unemployment are steady or ever-accelerating. Neither macro-economic policy makers, nor the elected officials and electorates to whom they are responsible, can avoid weighing the costs of unemployment against those of inflation. As Phelps has pointed out, this social choice has an intertemporal dimension. The social costs of involuntary unemployment are mostly obvious and immediate. The social costs of inflation come later.

What are they? Economists' answers have been remarkably vague, even though the prestige of the profession has reinforced the popular view that inflation leads ultimately to catastrophe. Here indeed is a case where abstract economic theory has a powerful hold on public opinion and policy. The prediction that at low unemployment rates inflation will accelerate toward ultimate disaster is a theoretical deduction with little empirical support. In fact the weight of econometric evidence has been against acceleration, let alone disaster. Yet the deduction has

been convincing enough to persuade this country to give up billions of dollars of annual output and to impose sweeping legal controls on prices and wages. Seldom has a society made such large immediate and tangible sacrifices to avert an ill defined, uncertain, eventual evil.

According to economic theory, the ultimate social cost of anticipated inflation is the wasteful use of resources to economize holdings of currency and other noninterest-bearing means of payment. I suspect that intelligent laymen would be utterly astounded if they realized that *this* is the great evil economists are talking about. They have imagined a much more devastating cataclysm, with Vesuvius vengefully punishing the sinners below. Extra trips between savings banks and commercial banks? What an anti-climax!

With means of payment—currency plus demand deposits—equal currently to 20 percent of *GNP*, an extra percentage point of anticipated inflation embodied in nominal interest rates produces in principle a social cost of 2/10 of 1 percent of *GNP* per year. This is an outside estimate. An unknown, but substantial, share of the stock of money belongs to holders who are not trying to economize cash balances and are not near any margin where they would be induced to spend resources for this purpose. These include hoarders of large denomination currency, about one-third of the total currency in public hands, for reasons of privacy, tax evasion, or illegal activity. They include tradesmen and consumers whose working balances turn over too rapidly or are too small to justify any effort to invest them in interest-bearing assets. They include corporations who, once they have been induced to undertake the fixed costs of a sharp-pencil money management department, are already minimizing their cash holdings. They include businessmen who are in fact being paid interest on demand deposits,

although it takes the form of preferential access to credit and other bank services. But, in case anyone still regards the waste of resources in unnecessary transactions between money and interest-bearing financial assets as one of the major economic problems of the day, there is a simple and straightforward remedy, the payment of interest on demand deposits and possibly, with ingenuity, on currency too.

The ultimate disaster of inflation would be the breakdown of the monetary payments system, necessitating a currency reform. Such episodes have almost invariably resulted from real economic catastrophes—wars, defeats, revolutions, reparations—not from the mechanisms of wage-price push with which we are concerned. Acceleration is a scare word, conveying the image of a rush into hyperinflation as relentlessly deterministic and monotonic as the motion of falling bodies. Realistic attention to the disaggregated and stochastic nature of wage and price movements suggests that they will show diverse and irregular fluctuations around trends that are difficult to discern and extrapolate. The central trends, history suggests, can accelerate for a long, long time without generating hyper-inflations destructive of the payments mechanism.

Unanticipated inflation, it is contended, leads to mistaken estimates of relative prices and consequently to misallocations of resources. An example we have already discussed is the alleged misallocation of time by workers who over-estimate their real wages. The same error would lead to a general over-supply by sellers who contract for future deliveries without taking correct account of the increasing prices of the things they must buy in order to fulfill the contract. Unanticipated deflation would cause similar miscalculations and misallocations. Indeed, people can make these same mistakes about relative prices even when the price level is stable. The mistakes are more likely, or the more

costly to avoid, the greater the inflationary trend. There are costs in setting and announcing new prices. In an inflationary environment price changes must be made more frequently—a new catalog twice a year instead of one, or some formula for automatic escalation of announced prices. Otherwise, with the interval between announcements unchanged, the average misalignment of relative prices will be larger the faster the inflation. The same problem would arise with rapid deflation.

Unanticipated inflation and deflation—and unanticipated changes in relative prices—are also sources of transfers of wealth. I will not review here the rich and growing empirical literature on this subject. Facile generalizations about the progressivity or equity of inflationary transfers are hazardous; certainly inflation does not merit the cliché that it is “the cruelest tax.” Let us not forget that unemployment has distributional effects as well as dead-weight losses.

Some moralists take the view that the government has promised to maintain the purchasing power of its currency, but this promise is their inference rather than any pledge written on dollar bills or in the Constitution. Some believe so strongly in this implicit contract that they are willing to suspend actual contracts in the name of anti-inflation.

I have long contended that the government should make low-interest bonds of guaranteed purchasing power available for savers and pension funds who wish to avoid the risks of unforeseen inflation. The common objection to escalated bonds is that they would diminish the built-in stability of the system. The stability in question refers to the effects on aggregate real demand, *ceteris paribus*, of a change in the price level. The Pigou effect tells us that government bondholders whose wealth is diminished by inflation will spend less. This brake on old-fashioned *gan*

inflation will be thrown away if the bonds are escalated. These considerations are only remotely related to the mechanisms of wage and price inflation we have been discussing. In the 1970's we know that the government can, if it wishes, control aggregate demand—at any rate, its ability to do so is only trivially affected by the presence or absence of Pigou effects on part of the government debt.

In considering the intertemporal tradeoff, we have no license to assume that the natural rate of unemployment is independent of the history of actual unemployment. Students of human capital have been arguing convincingly that earning capacity, indeed transferable earning capacity, depends on experience as well as formal education. Labor markets soggy enough to maintain price stability may increase the number of would-be workers who lack the experience to fit them for jobs that become vacant.

Macro-economic policies, monetary and fiscal, are incapable of realizing society's unemployment and inflation goals simultaneously. This dismal fact has long stimulated a search for third instruments to do the job: guideposts and incomes policies, on the one hand, labor market and manpower policies, on the other. Ten to fifteen years ago great hopes were held for both. The Commission on Money and Credit in 1961, pp. 39–40, hailed manpower policies as the new instrument that would overcome the unemployment-inflation dilemma. Such advice was taken seriously in Washington, and an unprecedented spurt in manpower programs took place in the 1960's. The Council of Economic Advisers set forth wage and price guideposts in 1961–62 in the hope of "talking down" the Phillips curve (pp. 185–90). It is discouraging to find that these efforts did not keep the problem of inflationary bias from becoming worse than ever.

So it is not with great confidence or optimism that one suggests measures to

mitigate the tradeoff. But some proposals follow naturally from the analysis, and some are desirable in themselves anyway.

First, guideposts do not wholly deserve the scorn that "toothless jawboning" often attracts. There is an arbitrary, imitative component in wage settlements, and maybe it can be influenced by national standards.

Second, it is important to create jobs for those unemployed and discouraged workers who have extremely low probability of meeting normal job specifications. Their unemployment does little to discipline wage increases, but reinforces their deprivation of human capital and their other disadvantages in job markets. The National Commission on Technology, Automation and Economic Progress pointed out in 1966 the need for public service jobs tailored to disadvantaged workers. They should not be "last resort" or make-work jobs, but regular permanent jobs capable of conveying useful experience and inducing reliable work habits. Assuming that the additional services produced by the employing institutions are of social utility, it may well be preferable to employ disadvantaged workers directly rather than to pump up aggregate demand until they reach the head of the queue.

Third, a number of measures could be taken to make markets more responsive to excess supplies. This is the kernel of truth in the market-power explanation of inflationary bias. In many cases, government regulations themselves support prices and wages against competition. Agricultural prices and construction wages are well-known examples. Some trade unions follow wage policies that take little or no account of the interests of less senior members and of potential members. Since unions operate with federal sanction and protection, perhaps some means can be found to insure that their memberships are open and that their policies are responsive to the unemployed as well as the employed.

As for macro-economic policy, I have

argued that it should aim for unemployment lower than the zero-inflation rate. How much lower? Low enough to equate unemployment and vacancies? We cannot say. In the nature of the case there is no simple formula—conceptual, much less statistical—for full employment. Society cannot escape very difficult political and intertemporal choices. We economists can illuminate these choices as we learn more about labor markets, mobility, and search, and more about the social and distributive costs of both unemployment and inflation. Thirty-five years after Keynes, welfare macroeconomics is still a relevant and challenging subject. I dare to believe it has a bright future.

## REFERENCES

- W. H. Beveridge, *Full Employment in a Free Society*, New York 1945.
- P. Doeringer and M. Piore, *Internal Labor Markets and Manpower Analysis*, Lexington, Mass. 1971.
- M. Friedman, "The Role of Monetary Policy," *Amer. Econ Rev.*, Mar. 1968, 58, 1–17.
- R. Hall, "Why is the Unemployment Rate so High at Full Employment?," *Brookings Papers on Economic Activity*, 3, 1970, 369–402.
- J. M. Keynes, *The General Theory of Employment, Interest, and Money*, New York 1936.
- A. Leijonhufvud, *On Keynesian Economics and the Economics of Keynes*. New York 1968.
- R. G. Lipsey, "The Relation Between Unemployment and the Rate of Change of Money Wage Rates in the United Kingdom, 1862–1957: A Further Analysis," *Economica*, Feb. 1960, 27, 1–31.
- H. S. Parnes, *Research on Labor Mobility*, Social Science Research Council, Bull. 65, New York 1954.
- G. L. Perry, "Changing Labor Markets and Inflation," *Brookings Papers on Economic Activity*, 3, 1970, 411–41.
- E. S. Phelps et al., *Micro-economic Foundations of Employment and Inflation Theory*, New York 1970.
- A. W. Phillips, "The Relation Between Unemployment and the Rate of Change of Money Wage Rates in the United Kingdom, 1861–1957," *Economica*, Nov. 1958, 25, 283–99.
- L. G. Reynolds, *The Structure of Labor Markets*, New York 1951.
- C. L. Schultze, "Has the Phillips Curve Shifted? Some Additional Evidence," *Brookings Papers on Economic Activity*, 2, 1971, 452–67.
- J. Tobin, "A Note on the Money Wage Problem," *Quart. J. Econ.*, May 1941, 55, 508–16.
- Commission on Money and Credit, *Money and Credit: Their Influence on Jobs, Prices, and Growth*, Englewood Cliffs 1961.
- Economic Report of the President 1962*, Washington 1962.
- U.S. National Commission on Technology, Automation, and Economic Progress, *Technology and the American Economy*, Washington 1966.

# Private and Social Rates of Return to Education of Academicians

By DUNCAN BAILEY AND CHARLES SCHOTTA\*

This paper presents the results of our examination of evidence on cost and returns to graduate education in the United States. The basic data include salary of all faculty members in the 829 highest paying U.S. colleges and universities during academic year 1966. The data base is all faculty at the rank of assistant professor or above at these institutions. The direct costs which are used to secure the net benefit stream for these faculty is the average cost structure for the production of Ph.D. degrees in all disciplines at the Berkeley and Los Angeles campuses of the University of California for the academic year 1966.

One large element of cost in producing a graduate degree is the income lost while attending graduate school. We have estimated this opportunity cost through the use of actual salaries prevailing in occupations open only to holders of the bachelor's degree in California in the fiscal year 1966.

Although not intended to be a complete listing of the limitations of this study, we should nevertheless like to point out that our results and measurements apply only to graduate degree holders who work for one of the 829 highest paying colleges and

universities in the United States. We have not examined the rate of return to graduate education for persons employed in governmental agencies or business organizations. Another limitation on our results relates to the comprehensiveness of our measure of social costs and benefits of graduate education. Our measures assume that graduate education in the United States occurs purely for investment reasons. Therefore we have not included intangible benefits to either society or the individual graduate, arising from the existence of a body of technically trained persons. If one believes that the existence of a body of people with graduate technical training in physics, biochemistry, statistics, French, or any other "university graduate specialty" confers "external economies," then one must also believe that the same kinds of benefits are conferred upon society by the existence of technically trained physicians, lawyers, bricklayers, plumbers, and automobile mechanics. We are not prepared, in the absence of any data on these measurements, to argue anything other than a test-of-the-market-place conclusion.

## I. Private Returns

In order to estimate the return to graduate education it is first necessary for the investigator to secure an estimate of the earnings of individuals in various occupations, which would be secured over a lifetime through the use of the human capital created during the course of securing a bachelor of arts degree. This lifetime stream represents the basic opportunity cost of not entering the job market with a

\* Senior economist, Business Economics Groups, W. R. Grace & Company, and associate professor of economics, Virginia Polytechnic Institute and State University, respectively. Bailey was instructor in economics at VPI&SU when the paper was prepared. An earlier version of this paper was presented at the Western Economic Association meetings, Davis, California, August 28, 1970. Our thanks must go to Vic Bonomo, James Buchanan, Paul Hamelman, Wilson Schmidt, Norman Starler, and an anonymous referee for helpful comments on earlier drafts of this paper. We are responsible for any taint which may remain in our results and conclusions.

bachelor's degree but instead utilizing that time for an investment in the acquisition of additional human capital. This basic income stream also represents the standard against which any increment (or decrement) that results from the investment in graduate education must be measured in order to assess the rate of return to this further human capital investment.

The basic income stream which we utilize is derived from State of California statistics. Each year the California State Personnel Board, in conjunction with local personnel agencies, conducts a salary survey in order to determine competitive salaries for state agency personnel in various categories. In the spring of 1967 such a survey was conducted and a basic sample of 860,000 employees was obtained. Working with the California State Personnel Board we were able to separate out those employees working at jobs requiring a four-year bachelor of science degree in California. Additionally, using State of California promotion and salary patterns we were able to construct a lifetime stream of earnings for the average holder of a four-year college or university degree. A sample-size weighted average across occupations was used in order to obtain the base income series for this study.<sup>1</sup>

The next step is to secure a similar income series for those who go on to graduate study. Our procedure can be set forth very simply. As a first step we examined the distribution of occupations of doctoral degree recipients in the United States by employer type. These data are summarized and presented in Table I. We note that 73 percent of those individuals who received doctoral degrees in the United States during years 1958-66 accepted academic employment with an ad-

ditional 13 percent working for governmental and nonprofit organizations, leaving only 14 percent in business organizations. When one examines the distribution, by employer type, of specific areas within all academic concentrations one is struck by the fact that almost all recipients of degrees in the arts and humanities work in academic employments. A close second, in terms of percentage recipients working in academe, is the field of education, followed by social sciences and biological sciences, each with slightly less than three-fourths of the recipients working in academe. The physical sciences trail with approximately 51 percent of its degree recipients in academic pursuits.

TABLE 1—OCCUPATIONS OF DOCTORAL RECIPIENTS IN THE UNITED STATES BY EMPLOYER TYPE, 1958-66  
(in percentages)

	Academic	Government and Nonprofit	Business
All	73	13	14
Arts & Humanities	96	3	1
Social Sciences	72	24	4
Physical Sciences	51	11	38
Biological Sciences	71	19	10
Education	91	8	1

*Source: Doctorate Recipients From United States Universities, 1958-1966*

From a market price standpoint the salary schedules in academic employment and government and nonprofit employment tend to be tied closely. It is frequently alleged and available data seem to confirm that holders of graduate degrees in business pursuits receive salaries appreciably above those in academic, government, and nonprofit occupations. When government and nonprofit are collapsed together with academic employments, it can be seen that 86 percent of all doctoral recipients work for academic institutions or in occupations where the salary struc-

<sup>1</sup> A detailed explanation of the construction of this base income series can be found in Bailey and Schotta (1969).

ture can be expected to be extremely close to those academically. Thus only 14 percent of the individuals who constitute our probable sample can be clearly excepted from the general conclusions of our study. Our results apply to at least 73 percent of those individuals investing successfully in graduate education with the percentage possibly running as high as 86 percent.

As the next step in securing a base salary stream for holders of the graduate degree working in academic (or closely related) pursuits, we have selected the 1966 published salary scale and published usual promotion schedule at the University of California (U.C.) campuses as our benchmark. Procedurally we have postulated a hypothetical holder of a graduate degree who is initially employed by any one of the U.C. campuses as an assistant professor, step II, at the salary prevailing for that position in academic year 1966. The U.C. salary and promotion schedule shows that the average time spent in each of the assistant professor steps (except assistant professor, step IV) is two years. The same normal time in grade prevails in the three associate professor ranks with three years being the usual salary period for full professor, steps I and II.

We have assumed that our faculty member was not distinguished by a performance which would warrant faster promotion than the average, nor was his career marred by performance which caused a slower than normal promotion rate. By selecting the average promotion pattern and applying it to the 1966 salary scale, we have secured for each year of his working life a salary and rank for this hypothetical University of California faculty member. We have chosen to conclude his career in the full professor, step III position since relatively few U.C. faculty members ever are promoted to the full professor, step IV or step V ranks. This, then, is our base salary stream for the U.C.

typical faculty member in 1966 salary and promotion terms.<sup>2</sup>

The next step was to convert this reference income stream into income streams for faculty in *U.S.* colleges and universities at large. The American Association of University Professors (*AAUP*) annually conducts a salary survey based upon data reported by institutions of higher education by professorial ranks. The *AAUP* has established ratings for overall faculty salary and ratings for faculty salaries by academic ranks from professor to assistant professor. Only one tenth of one percent of *U.S.* faculty in 1966 worked in universities with the highest, or AA salary rating; 3.4 percent worked in institutions with the A rating; almost 17 percent worked in institutions with the B rating; and fully 70 percent of the staff of *U.S.* universities were in institutions with the C and D ratings. Lower ratings cover 10 percent of the *U.S.* faculty.

TABLE 2—DISTRIBUTION OF PROFESSORIAL RANK FACULTY IN *U.S.* COLLEGES AND UNIVERSITIES AMONG *AAUP* AVERAGE SALARY RATINGS BY RANK IN 1966

Pattern	Asst Prof	Assoc Prof	Full Prof	Percent Total Sample
1	AA	AA	AA	.06 ( .1)
2	AA	AA	A	2.66 ( 3.8)
3	AA	A	A	2.28 ( 3.3)
4	A	A	A	4.1 ( 5.9)
5	AA	A	B	2.0 ( 2.9)
6	A	A	B	13.8 (19.8)
7	A	B	B	12.6 (18.1)
8	B	B	B	4.9 ( 7.0)
9	A	B	C	8.6 (12.3)
10	B	B	C	18.8 (26.9)
				69.66 100.0

Source: *AAUP* Bulletin

Table 2 presents the 1966 distribution of professorial rank faculty in *U.S.* colleges

<sup>2</sup> The data on average salaries and average promotion times at the University of California are from the California *Budget Supplement*.

203417

and universities among the *AAUP* average salary grades for these ranks. We have not considered for the purposes of our study faculty members working in universities with salary patterns lower than a BBC pattern.

We have delineated the ten highest patterns which prevail in universities. Pattern four is a university such that all three professorial ranks have a rating of A. In 1966 this was the University of California situation. The salary pattern for our hypothetical member therefore corresponds to salary pattern number four. As Table 2 indicates, in 1966, 4.1 percent of U.S. faculty members worked in universities with such a salary pattern. Taking this normalized U.C. salary, year by year, as the mid-point of pattern four, we calculated the percentage relation between the A rated salary for assistant professor in 1966 and the AA, B, and C ratings. Utilizing the same calculations for associate professor and full professor it was possible for us to secure an annual salary for faculty members at each of the salary patterns. For example, it was possible not only to secure the annual income for our hypothetical faculty member who went to work for U.C. and progressed through the ranks, it was also possible for us to secure an estimate for a young faculty member who began his career at any institution which paid A, remained at that same institution for an A associate professor and an A full professor. Likewise it was possible to estimate the income stream for the AA assistant professor, A associate Professor, A full professor or any other of the salary patterns.<sup>3</sup> We secured an eleventh salary stream which was the weighted average

annual income, term by term, of faculty members where the separate patterns one through ten were combined, with the weights in securing the average being the percentage of total faculty in 1966 in the 829 colleges and universities which occupied that salary pattern.

The next step was to secure the salary differential stream between each of the eleven faculty salary patterns and the undergraduate base income stream previously described.

In our analysis we do not consider a graduate school career toward the Ph.D. shorter than two years nor longer than six years. Therefore we have prepared the salary differential streams for opportunity cost purposes, for five alternate assumptions with respect to years in graduate school. As an example, our first assumption is that the faculty members in all eleven salary patterns spent two years in graduate school. We have used \$2500 as the income for the graduate earning stream in these two years. We believe this figure is a reasonable estimate of the average earning from teaching and fellowship employment in 1966 for the average graduate student.<sup>4</sup> This figure of \$2500 was subtracted from the average B.A. earnings for years one and two subsequent to graduation at the baccalaureate level to secure the opportunity costs of foregoing B.A. employment and continuing in graduate school.

We also prepared a salary stream for each of the eleven patterns under the assumption that the first three salary terms for every graduate degree salary pattern was \$2500. The same manipulation was undertaken for the first four years, the

<sup>3</sup> It is not necessary to assume that a given faculty member remains at the same university his entire working life. In order to achieve pattern four, for instance, it is only necessary that the faculty member be continuously employed at an institution which pays an A level salary for the rank he then holds.

<sup>4</sup> We assume that this market price is equal to the value of the marginal product of the graduate student in this use. To the extent that the graduate student receives a teaching assistantship and is actually *underpaid* for services provided, he reduces the costs of producing B.A.s. A model which incorporates this feedback effect of factor prices on output of the higher education business is described in Bailey and Schotta (1971).

first five years, and the first six years. Thus we emerged with fifty-five differential salary streams, eleven for each assumption with respect to the number of years in graduate school.

For our analysis we made seven basic assumptions. The first assumption is that we employ constant dollar amounts, year by year, in order to eliminate the effect of inflation. This directly gives us rates of return expressed in real terms.

Second, in constructing our time streams of income, we assume that there is no relative change in the structure of supply and demand in the labor market which would cause incomes of baccalaureate-degree holders to rise relative to incomes of graduate-degree holders or vice versa.<sup>5</sup>

Third, we assume in our model that higher education is viewed by the state as an investment in human capital to the complete exclusion of the production of any consumer good component. Thus we eliminate, in the publicly funded university context, the necessity for evaluating higher education in any consumer goods setting.

Fourth, we assume that the existence of this higher-education-produced human capital is of the same general type as most "producer's goods." This is equivalent to stating that society, in terms of externalities, is just as well off proportionately by the production of the human capital embodied in an engineer as it is in production of an additional unit of physical capital.

Fifth, we assume that the wage of a degree holder is approximately equal to the

rent on his human capital. This simply denies any significant raw labor element in the salary of a college graduate or Ph.D. recipient.

Sixth, we assume that the average product of human capital is equal to the marginal product of human capital, an assumption quite common to capital theory. By this assumption we avoid some of the Marxian interpretations of the labor market for human capital.

Seventh, we assume that the structure of the market for human capital output is imperfectly competitive and is characterized by monopolistic and monopsonistic competition.

The raw differential salary streams so far obtained are not useful for any investment analysis of graduate education. We have, therefore, made use of an algorithm which finds the discount rate which equalizes the present values of the two comparative streams, the base income stream,  $B$ , and one of the income patterns under one of the alternate assumptions with respect to years spent in graduate school in the educational investment process,  $G$ . Equation (1) presents the algorithm in compact notation. We have utilized a stopping rule such that we are seeking the interest rate,  $r_{jv}$ , for each set of  $j$  salary patterns where  $j$  equals one through eleven; and  $v$  years in graduate school where  $v$  equals two through six, such that the present value of the differential income stream for each year is less than or equal to the absolute value of \$10.<sup>6</sup>

<sup>5</sup> Although there may well be relative changes in the two markets through time, we have no a priori reason to assume that Ph.D. holders will fare better (or worse) than B.A. or B.S. degree holders. The year 1966 is a particularly good base since there was relatively full employment in both markets. The experience since 1966 suggests that the Ph.D. holder may be worse off (relative to the B.A.) than he was in 1966, but this effect may be transitory only. If it persists, then the rates of return presented in this paper will be overstated.

<sup>6</sup> The internal rate of return will be unique and unbiased as long as the terms in the net returns stream do not cycle about the abscissa (a semi-convex set). By inspection we have verified the semi-convexity of all fifty-five net returns streams we work with. For a complete discussion of these conditions, see Jack Hirshleifer. Since we employ an average investment in human capital with a single duration and do not consider the possibility of truncation of this average duration (e.g., changes in mandatory retirement ages), the condition for a unique internal rate of return discussed by Kenneth Arrow and David Levhari, and amplified by C. (over)

$$(1) \quad |\$10| \geq \sum_{i=1}^{43} \frac{G_{ij} - B_i}{(1+r_{jv})^i} \quad j=1, 2, \dots, 11; \\ v=2, 3, \dots, 6$$

When equation (1) is applied to the data, we secure the internal rates of return for the various salary patterns based on alternate assumptions with respect to years in graduate school. Table 3 shows the internal rate of return to an investment in graduate education for each of eleven salary patterns and five graduate school residence periods. The extreme importance of the number of years spent in the educational investment process on internal rates of return by years in graduate school is shown in Table 3.

Although the National Academy of Sciences indicates that 5.1 years is the average length of time a graduate student is *registered* for the Ph.D. degree, we have selected four years as the average length of time physically spent in graduate school (and therefore out of the academic job market) by the average successful graduate student.

Hence we have chosen to focus our discussion on the internal rates of return for each of the salary patterns and for the average pattern for this number of years spent in graduate school. These internal rates of return for education beyond the B.A. degree range from a high of 11.6 percent for pattern one to zero or negative returns for patterns nine and ten.<sup>7</sup> Earlier in Table 2, we noted that slightly more than a quarter of the faculty in these colleges and universities were in salary patterns six and seven. Table 3 indicates that these faculty members achieve an

TABLE 3—INTERNAL RATE OF RETURN FOR THE  
AAUP SALARY PATTERNS BASED ON ALTERNATE  
ASSUMPTIONS WITH RESPECT TO  
YEARS IN GRADUATE SCHOOL

Pattern	Years in Graduate School				
	2	3	4	5	6
1	20.3	14.7	11.6	9.6	8.0
2	17.8	12.1	9.1	7.1	5.6
3	15.4	10.6	8.0	6.2	4.9
4	12.3	9.0	7.0	5.5	4.3
5	10.9	6.2	3.7	2.0	0.7
6	7.5	4.5	2.6	1.2	0.0
7	6.6	4.0	2.3	1.0	—
8	5.3	3.2	1.7	0.5	—
9	—	—	—	—	—
10	—	—	—	—	—
11	4.9	2.5	0.8	—	—

average rate of return to graduate school of around 2.4 percent. Another 25 percent of faculty who are in patterns nine and ten have zero or negative rates of return.

Pattern eleven, which is a weighted average salary, shows less than one percent internal rate of return for four years in graduate school.<sup>8</sup>

The academic year salary paid to university personnel generally measures the market valuation of their human capital services to the state for the teaching, research, and service function. Many faculty members, however, work during the balance of the calendar year; that is, after the academic year is concluded they may teach, or they may be paid throughout the entire year in some outside consulting activity. Unfortunately, information on the percentage of faculty members who teach every summer of their working life or who engage in extensive consulting activities each year is fragmentary. Since we cannot measure this additional income directly, we have calculated the percentage of aca-

Norström does not apply to our calculated rates of return. However, for individual investment decisions with the corresponding possibilities of early investment truncation, their work would be directly applicable.

<sup>7</sup> By a negative return we mean that the undiscounted sum of the net returns stream is negative. Since such a stream will yield only imaginary roots, we cannot discuss a true negative rate of return.

<sup>8</sup> Orley Ashenfelter and James Mooney, when comparing lifetime income streams of Ph.D. recipients who held Woodrow Wilson fellowships to those of Woodrow Wilson fellowship holders who *quit* graduate school, found a rate of return in excess of 5 percent. We are not certain to what this rate should apply.

demic year salary which would have to be earned each year to raise the internal rates of return presented in Table 3 to certain target rates of return.

Below we present equation (2) which has as its objective evaluating the expression on the right-hand of the equality/inequality sign for each interest rate between 2 and 10 percent, such that  $\lambda_{jv}$  is found for each of the  $j$  salary patterns for each of the  $v$  years in graduate school, such that equation (2) is satisfied. Again the stopping rule for present value items is the absolute value of \$10.

$$(2) \quad |S10| \geq \sum_{i=1}^{43} \frac{(\lambda_{jv} G_{ij}) - B_i}{(1+r^i)}$$

$$j = 1, 2, \dots, 11;$$

$$v = 2, 3, \dots, 6;$$

$$100r = 2, 3, \dots, 10$$

Procedurally, the algorithm takes each of the fifty-five graduate earning schemes and for a fixed discount rate finds a multiplier  $\lambda$  such that the stopping rule is met when the annual graduate income is multiplied by  $\lambda$  and the resultant graduate-baccalaureate salary differential is summed over the forty-three working years subsequent to graduation with a baccalaureate degree. The results of this set of calculations are presented in Table 4 for the faculty member who spends four years in graduate school.

Referring back to Table 3 we can see that for salary pattern one the internal rate of return is in excess of 6 percent, therefore the  $\lambda$  necessary to produce a 6 percent rate of return was one or less than one. However, for salary pattern five the internal rate of return for the four-year Ph.D. was 3.7 percent. The answer to the question of what percentage of the academic year's salary a typical individual in salary pattern five would have to earn each year, every year of his working life, in

TABLE 4—PERCENT OF ACADEMIC YEAR SALARY NECESSARY TO SECURE STATED INTERNAL RATE OF RETURN FOR THE FOUR YEAR PH.D.

Pattern	Rate of Return				
	2	5	6	7	10
1					
5		2.6	4.8	7.1	14.3
9	14.5	20.1	22.2	24.4	31.3
10	16.3	22.9	25.3	27.9	36.0
11	2.4	9.5	12.2	14.9	23.5

order to raise his internal rate of return for the investment in graduate education to 6 percent, is 4.8 percent. Simply stated, in order to give a 6 percent rate of return on the investment in graduate education for the four-year Ph.D. working at salary pattern five, this individual would have to earn an amount over and above his academic year salary, each and every year of his working life equal to approximately 5 percent of whatever his annual salary happened to be in each year.

When one examines salary pattern eleven, which is the weighted average academic year salary for the four-year Ph.D., we see that an amount equal to 12 percent of the academic year salary would have to be earned each and every year to raise the internal rate of return for all Ph.D. holders working in universities to a target rate of return of 6 percent.

With the presentation of this data on incomes and internal rates of return for the investment in graduate education we have completed the development of private (real) rates of return to graduate education where the subsequent employment of human capital thus created is in the academic sector (and possibly the government and nonprofit sector). These are private (real) rates of return accruing to the individual before tax as a result of the investment in graduate education.

As the most likely estimate of the average private (real) rate of return to an in-

vestment in Ph.D. level graduate education, we advance the internal rate of return of .8 percent being earned by the weighted average salary pattern eleven for four years of time spent in graduate school.

## II. Social Returns

While in the previous section we calculated incomes and rates of return which represented private rates of return to individuals acquiring graduate education, we have not considered all of the costs of producing the increment in human capital as a result of graduate school attendance. The net income stream figures for the individual, while including the opportunity costs borne by the individual and an allowance for direct outlay which would not be made if the person were not in graduate school, do not include the social resources provided in most graduate education programs.

We should make one thing very clear before we begin our discussion of the social cost of providing the investment in graduate education. Our results and conclusions apply only to public universities funded principally through tax revenues. Universities such as the University of California and other major state universities are funded in their teaching and teaching-associated research activities through public tax revenues. Tuition charges are typically extremely low for in-state students and in general, graduate students who hold assistantships or fellowships are exempt from fees other than those charged state residents.

To secure cost figures for the production of graduate degrees we have the detailed budgetary allocation for the University of California for the academic year 1966. For budgeting purposes U.C. utilizes a weighting formula to estimate resource demands such that lower division undergraduate students have a weight of one, upper division undergraduate students have a

weight of one and one-half, M.A. and first-year graduate students have weight of two and one-half, and advanced graduate students have a weight of three.<sup>9</sup>

Through the use of this budgeting procedure resource requests for the University of California are developed. We have taken this procedure, taken the state-supplied current operating expense figure from the State of California *Budget Supplement*, allocated out the administrative loading to the various campuses and secured for each of the U.C. campuses an estimate of the average annual cost for first-year graduate education and advanced graduate education.

Because of the extremely high start-up costs associated with graduate education, we have excluded from our further analysis the average cost figures for all campuses except Berkeley and Los Angeles. This is necessitated by the fact that in the case of such campuses as San Diego or Davis, the average cost per graduate year was approximately two to three times the average

<sup>9</sup> For example, if there are 1,000 students evenly divided among the four utilization classes we have outlined, to secure the weighted allocation for resource budgeting purposes the 250 lower division undergraduates would be multiplied by one; the 250 upper division undergraduates would be multiplied by one and one-half; the 250 first-year graduate students would be multiplied by two and one-half, and the 250 advanced graduate students would be multiplied by three. The sum of this weighting would then equal 2,000, the weighted full time equivalent student population.

If a total budgetary allocation for these 1,000 students equaled \$100,000, \$12,500 of this \$100,000 would be allocated for lower division undergraduate instructional activities; \$18,750 would be allocated for upper division undergraduate instructional activities; \$31,250 for first-year graduate student instructional activities; and \$37,500 for advanced graduate instructional activities. To secure the budgeted average resource cost per individual per instructional level, one would simply divide these allocations per level by the actual students among whom these resources are to be divided. In the case of our example the average cost per lower division undergraduate student would be \$50. For the upper division undergraduate student it would be \$75. For the first-year graduate student the average cost would be \$125. For the advanced graduate student the average cost would be \$159.

of the two established campuses of the university.

In our further analysis we take the cost structure of these two campuses to be fairly representative of average cost encountered at major producers of graduate human capital in the United States. The average first-year resource cost which we have employed is \$4,358. For each advanced year the cost is \$6,035.

Since these resource outlays are made over a particular time span, ranging in our study from two to six years, clearly we are in a position to make an estimate of the present value cost of the resources publicly provided for the Ph.D. degree for the academic year 1966. Table 5 presents our calculations of this present value cost. For convenience, we have assumed, contrary to fact, that every student who enrolls in the first year of the Ph.D. program completes a Ph.D. degree in the specified number of years. If, as is well known in industrial quality and cost control, some defects are produced which are not subsequently sold and must be scrapped, then the resources expended on their production must be averaged over the acceptable units, thus increasing their average cost. As a result the present value costs presented in Table 5 must be considered as the lowest possible estimate which could be accepted.

TABLE 5—PRESENT VALUE COSTS OF A PH.D. DEGREE AT THE UNIVERSITY OF CALIFORNIA, BERKELEY, AND LOS ANGELES, 1966\*

Discount Rate (in percentage)	Years in Graduate School				
	2	3	4	5	6
5	\$10,104	\$15,574	\$20,787	\$25,752	\$30,481
6	10,049	15,420	20,487	24,867	29,376
7	9,996	15,267	20,194	24,798	29,101

\* Assumes no dropouts. Includes only current operating expense appropriation; does not include capital consumption allowance.

TABLE 6—PRESENT VALUE COST OF PH.D. DEGREES AT THE UNIVERSITY OF CALIFORNIA, BERKELEY, AND LOS ANGELES, 1966

Discount Rate (in percentage)	Years in Graduate School		
	3	4	5
5	24,176	29,389	34,354
6	23,969	29,036	33,816
7	23,765	28,692	33,296

Dropout assumption:

$X_1$  = total number entering graduate school  
 $X_2 = .75X_1$  = number entering year 2  
 $X_3 = .625X_1$  = number entering year 3  
 $X_4 = .5X_1$  = number entering year 4  
 $X_5 = .5X_1$  = number entering year 5

Includes only current operating expense appropriation; does not include capital consumption allowance.

Again selecting four years as the average length of time in graduate school which appears to be most likely we would like to focus on 6 percent as the appropriate social discount rate.<sup>10</sup> Here we can see that the average present value public cost of producing the Ph.D. degree at the University of California main campuses is around \$20,500. By viewing Table 5, one can see illustrated dramatically again the extreme importance in the cost structure of number of years spent in graduate school.

Since the dropout assumption is so crucial we have examined an alternative dropout assumption scheme which we believe is appropriately conservative. Table 6 presents the results of our calculations.

Again concentrating on 6 percent and four years in graduate school we see that the present value cost is \$29,000+. When

<sup>10</sup> In 1966, the implicit GNP deflator rose 2.71 percent, the consumer price index rose 2.91 percent, and the wholesale price index rose 3.32 percent. Taking these changes as rough approximations to the true rate of inflation, we can translate a 6 percent real rate of return into a money rate of return of between 8.71 and 9.32 percent. We have calculated our results for real rates of return between 2 and 10 percent, and for two through six years of graduate school. We will be happy to supply these tables upon request.

compared with the present value cost of 6 percent for four years of graduate school under the no dropout assumption of \$20,500 we see the extreme increase in the use of social resources which is produced by even a modest failure rate.

Against the background of this model of the graduate education investment process, let us turn to the results of our analyses of the social rate of return for the eleven salary patterns of the academic year for the alternate assumptions with respect to the number of years in graduate school.

Table 7 presents these results for the no dropout assumption. Since we have worked primarily with the four-year graduate school Ph.D., let us look at the social rate of return under the no dropout assumption which accrues to the investment in human capital for the various patterns.

TABLE 7—SOCIAL RATE OF RETURN FOR ACADEMIC YEAR AAUP SALARY PATTERNS, 1966, NO DROPOUT ASSUMPTIONS

Pattern	Years in Graduate School				
	2	3	4	5	6
1	14.0	10.1	7.8	6.2	5.0
2	11.4	7.6	5.4	3.9	2.7
3	10.1	6.8	4.8	3.4	2.3
4	8.7	6.0	4.3	3.0	1.9
5	5.8	2.7	0.9	—	—
6	4.3	1.9	0.3	—	—
7	3.9	1.7	0.1	—	—
8	3.2	1.2	—	—	—
9	—	—	—	—	—
10	—	—	—	—	—
11	3.5	1.4	—	—	—

Pattern one yields slightly less than 8 percent. Patterns two through four yield 4 to 6 percent, and all other patterns yield less than 1 percent.

The social return being earned on the investment in graduate education for pattern eleven is zero or negative.

In Table 8 we present the percent of academic year salary which would be necessary for the four-year Ph.D., with a

no dropout assumption, to earn, in addition to his academic salary, to secure various specified social rates of return. Selecting the 6 percent rate of return, we can see that pattern one already produces a social rate of return for the four-year Ph.D. in excess of 6 percent. For all other salary patterns, it is necessary to earn progressively larger percentages of the academic year's salary.

TABLE 8—PERCENT OF ACADEMIC YEAR SALARY NECESSARY TO SECURE STATED SOCIAL RATE OF RETURN FOR THE FOUR-YEAR PH.D., NO DROPOUT ASSUMPTIONS

Pattern	Rate of Return		
	5	6	7
1	—	—	—
5	15.2	20.0	25.3
9	37.6	43.2	49.5
11	23.6	29.5	36.8

The weighted average pattern, pattern eleven, for all four-year Ph.D.'s tells us that each faculty member working in an American college or university would have to earn up to an amount equal to almost 30 percent of his academic year's salary in addition to his academic salary each year of his working life to secure a 6 percent social rate of return to the investment in graduate education.

Table 9 presents the same data as Table

TABLE 9—SOCIAL RATE OF RETURN FOR ACADEMIC YEAR AAUP SALARY PATTERNS, 1966, DROP-OUT ASSUMPTIONS

Pattern	Years in Graduate School				
	2	3	4	5	6
1	11.8	8.7	6.9	5.6	4.5
2	9.3	6.3	4.6	3.3	2.2
3	8.2	5.6	4.0	2.8	1.8
4	7.2	5.0	3.5	2.4	1.5
5	4.1	1.7	0.2	—	—
6	3.1	1.0	—	—	—
7	2.8	0.8	—	—	—
8	2.2	0.5	—	—	—
9	—	—	—	—	—
10	—	—	—	—	—
11	1.4	—	—	—	—

7 with one difference: we have incorporated the drop-out scheme presented in detail above. The net effect of including this more realistic dropout assumption is to lower the social rates of return being earned in each pattern. The average pattern, pattern eleven, now shows a zero or negative social rate of return to investment in graduate education.

In Table 10 we can see, based on the dropout assumption scheme and based on the assumption that every faculty member spends four years in graduate school, it would be necessary for each and every faculty member to earn almost 39 percent of his academic year salary, in addition to the academic year salary, in order to secure a 6 percent social rate of return on the investment in graduate education.

TABLE 10—FOUR-PERCENT OF ACADEMIC YEAR SALARY NECESSARY TO SECURE STATED SOCIAL RATE OF RETURN FOR THE FOUR-YEAR PH.D. DROPOUT ASSUMPTIONS

Pattern	Rate of Return (in percentage)		
	5	6	7
1	—	—	0.6
5	21.3	27.7	34.5
9	46.4	54.3	63.4
11	31.0	38.7	47.5

Parenthetically, it might be noted that the best information we have been able to secure from informal discussion with a large number of educational administrators suggests that less than half of the university's faculty work at teaching in any given summer. Most educators estimate that less than half of their faculty do any significant amount of outside consulting. Our information suggests that in the arts and humanities, in general, a minimal amount of consulting on an outside basis is done. However, in engineering, the physical sciences, and certain of the social sciences, there is an appreciable amount of consulting work. The results of our inter-

views with university administrators suggest that no more than half of the university faculty members earn approximately 15 percent of their academic year's salary through outside activities each year. Certainly this estimate falls far short of the almost 39 percent which each faculty member would be required to earn each and every year of his working life to secure even the 6 percent social rate of return to investment in graduate education indicated in Table 10.

Let us now turn to the question of the profitability of investment in graduate education for society.

In Table 11 we present a summary set of figures which shows the cash value of the average excess above the academic year's salary which must be earned outside, each and every year by each faculty member, to achieve three stated social rates of return under the assumption that every faculty member spends four years in graduate school.

TABLE 11—CASH VALUE OF AVERAGE EXCESS OVER ACADEMIC YEAR SALARY WHICH MUST BE EARNED EACH YEAR TO ACHIEVE STATED SOCIAL RATE OF RETURN FOR FOUR-YEAR PH.D.

	5%	6%	7%
No Dropout	\$2,993	\$3,903	\$4,814
Dropout	\$4,033	\$5,035	\$6,180

Selecting 6 percent as a target rate of return we see that if we accept the no dropout assumption, faculty members must earn almost \$4,000 above their academic year salaries each and every year of their working life in order to raise the target rate of return to 6 percent. If we accept our modest dropout scheme, the required outside earning rises to slightly more than \$5,000 per year.

### III. Conclusions

1. The private real rate of return to

graduate education *in general* is either zero or less than 1 percent.

2. The social real rate of return to graduate education in general is either zero or less than 1 percent.<sup>11</sup>

3. In view of the findings of Becker, Bruce Wilkinson, and Bailey and Schotta (1969) of a private and social rate of return to investment in undergraduate education in excess of 10 percent, it seems likely that there should be a reallocation of educational expenditures by government from graduate education to undergraduate education for more efficient resource utilization, both socially and privately. However, the close relationship between the potential graduate degree holder and rates of return to investment in undergraduate education implies that this reallocation would lower the rate of return to investment in undergraduate education and raise the rate of return to investment in graduate education. For a discussion of the interrelationships between returns and costs in undergraduate and graduate education, see Bailey and Schotta (1971).

4. Although more specific measurements must be made of returns streams and costs, we hazard the guess that most of the resources to be removed from the investment process in graduate education should come from the humanities, education and certain of the social sciences for it is there that the returns streams appear low, relative to the bachelor's degree.<sup>12</sup>

<sup>11</sup> Although small, the exclusion of capital consumption allowances from our cost estimates tends to bias our social rate of return findings upward.

<sup>12</sup> We can only present rates of return for the average faculty member in the average discipline. In order to calculate rates of return by disciplines, it would be necessary to have disaggregated university salary streams and disaggregated B.A. salary streams, since it is likely that the higher paid disciplines are populated by individuals who would have obtained above average B.A. salaries. In the same manner, we are comparing average B.A.s with average Ph.D.s. Since salary patterns one through six are above the median income level of faculty salaries, the comparison of these patterns with the average B.A. income stream tends to *overstate* the rate of return to academics. Similarly, the rates of

5. Clearly, in addition to undertaking improvement and technological revisions in the human capital production process in graduate school to reduce the number of years spent in graduate school, great attention should be given to pre-selection and immediate quality control measures to reduce the dropout rate in graduate school. The dropout pattern appears, in terms of resource commitments, to be even more costly than the number of years spent in graduate school.

6. The returns to graduate education, rather than being pecuniary, are psychic. The positive capital sum loss which appears to be taken, on the average, by the graduate student (and most certainly by those in the "soft sciences" and humanities) is probably the "ticket" to a certain life style which has been described by the phrase "why work . . . get a Ph.D.!"<sup>13</sup>

#### REFERENCES

- K. Arrow and D. Levhari, "Uniqueness of the Internal Rate of Return with Variable Life of Investment," *Econ. J.*, Sept. 1969, 79, 560-66.
- O. Ashenfelter and J. D. Mooney, "Some Evidence on the Private Returns to Graduate Education," *Southern Econ. J.*, Jan. 1969, 35, 247-56.
- D. Bailey and C. Schotta, "The Private Rate

---

return shown for patterns eight through ten may be understated. Thus, all our conclusions must be based upon the comparison in pattern eleven, the weighted average pattern, between the *average* Ph.D. recipient and the *average* B.A. holder for which this bias does not exist.

<sup>13</sup> To the extent that a private consumption element to investment in graduate education in the humanities and "soft sciences" exists, a scheme of differential tuition charges would equalize social returns among fields and provide a market measure of the value of this "life style ticket." To the extent that this positive capital sum loss is the result of misinformation about relative streams (or mistaken expectations concerning salary level increases), such differential tuition schemes might merely introduce distortions in the resource allocation and occupational choice pattern. There appears to be no basis for selecting either of these two alternatives given our present empirical knowledge about such things as elasticity, relative costs among disciplines, etc.

- of Return to Investment in Undergraduate Education," *Proc. Bus. Econ. Statist. Sect., Amer. Statistical Ass.*, 1969, 460-67.
- and ———, "A Dynamic Programming Approach to Optimal Resource Allocation in Education," in B. Avi-Itzhak, ed., *Developments in Operations Research*, London 1971, 451-63.
- G. Becker, *Human Capital*, New York 1964.
- J. Hirshelifer, "On the Theory of Optimal Investment Decision," *J. Polit. Econ.*, Aug. 1958, 66, 329-52.
- C. Norström, "Uniqueness of the Internal Rate of Return with Variable Life of Investment: A Comment," *Econ. J.*, Dec. 1970, 80, 983-84.
- B. Wilkinson, "Present Values of Lifetime Earnings for Different Occupations," *J. Polit. Econ.*, Dec. 1966, 74, 556-73.
- Doctorate Recipients From United States Universities, 1958-66*, National Research Council-National Academy of Sciences, Washington 1967.
- "Further Progress: The Economic Status of the Profession," *AAUP Bull.*, June 1967, 53, 136-95.
- State of California, *Budget Supplement: Salaries and Wages, 1967*, Sacramento 1967.

# Distributional Equity and the Optimal Structure of Public Prices

By MARTIN S. FELDSTEIN\*

Several recent papers (William Baumol and David Bradford, Abba Lerner, Avinash Dixit) have restated the rules originally derived by Frank Ramsey and M. Boiteux for optimal pricing by a public enterprise that produces several goods and that must satisfy a budget constraint. Neither the Ramsey-Boiteux papers nor the more recent discussions deal directly with the important distributional aspects of public pricing. Ramsey actually developed his study as a derivation of the optimal excise taxes to be levied on a single individual. Boiteux specified his analysis to include an optimal lump sum income redistribution as well as the optimal pricing of publicly produced goods. The more recent papers have also limited their analysis to the derivation of prices that achieve Paretian efficiency.

In practice, optimal lump sum redistribution is impossible and the distributional aspect of public pricing is an important policy consideration. The Ramsey-Boiteux rule is therefore inadequate. The current paper extends these earlier results on public enterprise pricing by explicitly incorporating distributional equity.<sup>1</sup>

## I. The Distributional Characteristic and Optimal Pricing

It is sufficient to consider a public enterprise that produces two goods<sup>2</sup> and sells them at prices  $p_1$  and  $p_2$ . Let  $S(p_1, p_2, y)$

be the traditional consumer surplus of a household with income  $y$  that can purchase the goods at prices  $p_1$  and  $p_2$ . Let the distribution of household incomes be represented by the *relative* density function  $f(y)$ ; i.e., if there are  $N$  households in the population being served, the number in a small interval around  $y_0$  is  $Nf(y_0)dy$ . Finally, let the marginal social utility of a dollar to a household with income  $y$  be denoted  $u'(y)$ .<sup>3</sup> It will be assumed that the marginal utility of income to a household is unaffected by the prices charged by the public enterprise. This approximation seems justified for any practical application.<sup>4</sup>

The appropriate welfare maximand is the weighted sum of the household consumer surpluses, weighting by the marginal social utility of income to that household:<sup>5</sup>

$$(1) \quad W = N \int_0^\infty S(p_1, p_2, y) u'(y) f(y) dy$$

$W$  must be maximized subject to the con-

<sup>3</sup> The term "marginal social utility" is used here to denote the derivative of the social welfare function with respect to the income of the household. Since there are only a finite number of households, this derivative is well defined; the continuous density function  $f(y)$  and the associated integrals defined below should be regarded as approximations.

<sup>4</sup> Constant  $u'(y)$  for each household permits doing the analysis in terms of consumer surplus without specifying any particular Hicksian definition. It also makes it reasonable to assume that the argument of the marginal social utility function is money income, i.e., that  $u'(y)$  is not affected by changes in  $p_1$  and  $p_2$ . Since only the first-order conditions are relevant, the analysis can be developed without the use of consumer surplus by working with the indirect utility function; this approach is followed in my forthcoming article on the pricing of public intermediate goods.

<sup>5</sup> This assumes that for the other goods to which the marginal demand is transferred there is no producer surplus or excise tax revenue; see Dixit, A.C. Harberger, and Lerner.

\* Professor of economics, Harvard University.

<sup>1</sup> This question is also considered by Herbert Mohring and Peter Diamond and James Mirrlees. Mohring's equation (30), p. 703, and Diamond and Mirrlees' equation (65), p. 266, are equivalent alternative statements of the first-order optimality condition of equation (4) below.

<sup>2</sup> The problem of peak hour or peak season pricing is of course a special case of pricing different goods.

straint that revenue minus production cost be equal to a specified amount ( $B$ ). If  $q_i(p_1, p_2, y)$  is the quantity of good  $i$  purchased by a household with income  $y$ ,<sup>6</sup> the total quantity sold of good  $i$  is:

$$(2) \quad Q_i = N \int_0^\infty q_i(p_1, p_2, y) f(y) dy$$

Letting  $C(Q_1, Q_2)$  be the total production cost, the constrained maximand is the Lagrangian expression

$$(3) \quad Z = W + \lambda[p_1 Q_1 + p_2 Q_2 - C(Q_1, Q_2) - B]$$

To derive the first-order conditions for a maximum we make use of the result that  $\partial S(p_1, p_2, y)/\partial p_i = -q_i(y)$ , the quantity of good  $i$  purchased by a household with income  $y$ . The basic first-order conditions can be written:

$$(4a) \quad \frac{\partial Z}{\partial p_1} = -N \int_0^\infty q_1(y) u'(y) f(y) dy + \lambda \left[ Q_1 + p_1 \frac{\partial Q_1}{\partial p_1} + p_2 \frac{\partial Q_2}{\partial p_1} - m_1 \frac{\partial Q_1}{\partial p_1} - m_2 \frac{\partial Q_2}{\partial p_1} \right] = 0$$

$$(4b) \quad \frac{\partial Z}{\partial p_2} = -N \int_0^\infty q_2(y) u'(y) f(y) dy + \lambda \left[ p_1 \frac{\partial Q_1}{\partial p_2} + Q_2 + p_2 \frac{\partial Q_2}{\partial p_2} - m_1 \frac{\partial Q_1}{\partial p_2} - m_2 \frac{\partial Q_2}{\partial p_2} \right] = 0$$

where  $m_i = \partial C/\partial Q_i$ , the marginal cost of good  $i$ .

A convenient concept for introducing considerations of distributional equity in the analysis of optimal prices and taxes is the *distributional characteristic* of a good.

The distributional characteristic of good  $i$  is defined by the ratio

$$(5) \quad R_i = \frac{N}{Q_i} \int_0^\infty q_i(y) u'(y) f(y) dy$$

The ratio  $R_i$  is a weighted average of the marginal social utilities, each household's marginal social utility weighted by that household's consumption of good  $i$ . The conventional welfare assumption that  $u'(y)$  declines as  $y$  increases implies that the value of  $R_i$  will be greater for a necessity than for a luxury. The higher the income elasticity of demand for a good, the lower the value of  $R_i$ . The next section illustrates an approach to making  $R_i$  an operational measure. First, however, we derive a number of results that do not depend on any specific parametric representation of  $R_i$ .

Equations (4a) and (4b) can be expressed in terms of the  $R_i$ 's and simplified by denoting the price elasticity as

$$\epsilon_{ij} = \frac{\partial Q_i}{\partial p_j} \frac{p_j}{Q_i}$$

and by employing the Slutsky relation  $\epsilon_{ij} = \epsilon_{ji} p_j Q_j / p_i Q_i$ .<sup>7</sup> The first-order conditions are then:

$$(6a) \quad R_1 = \lambda \left[ 1 + \epsilon_{11} \left( \frac{p_1 - m_1}{p_1} \right) + \epsilon_{12} \left( \frac{p_2 - m_2}{p_2} \right) \right]$$

$$(6b) \quad R_2 = \lambda \left[ 1 + \epsilon_{21} \left( \frac{p_1 - m_1}{p_1} \right) + \epsilon_{22} \left( \frac{p_2 - m_2}{p_2} \right) \right]$$

<sup>7</sup> The use of the Slutsky compensated demand relation  $\epsilon_{ij} = \epsilon_{ji} p_j Q_j / p_i Q_i$ , ignores the income effect. If this is taken into account, the quantity  $\alpha_1 S_1 (R_2 - \lambda) - \alpha_2 S_2 (R_1 - \lambda)$ , where  $\alpha_i$  is the income elasticity of demand and  $S_i$  is the share of income spent on good  $i$ , must be added to the numerator and subtracted from the denominator of the right-hand side of equation (7). Since  $S_1$  and  $S_2$  can be expected to be quite small and the weighted difference even smaller, this correction for the income effect is likely to be of no practical significance.

<sup>6</sup> This implicitly assumes that all other prices remain constant.

These may be solved explicitly to yield the relative "profit" or "tax" rates:

$$(7) \quad \frac{(p_1 - m_1)/p_1}{(p_2 - m_2)/p_2} = \frac{\epsilon_{22}(R_1 - \lambda) - \epsilon_{12}(R_2 - \lambda)}{\epsilon_{11}(R_2 - \lambda) - \epsilon_{21}(R_1 - \lambda)}$$

In the special case in which the distributional characteristics are irrelevant, i.e.,  $R_1 = R_2$ , equation (7) yields the basic Ramsey rule:

$$(8) \quad \frac{(p_1 - m_1)/p_1}{(p_2 - m_2)/p_2} = \frac{\epsilon_{22} - \epsilon_{12}}{\epsilon_{11} - \epsilon_{21}}$$

If, moreover, the cross price elasticities are zero ( $\epsilon_{21} = \epsilon_{12} = 0$ ), we obtain the familiar rule that the tax rates should be inversely proportional to the own price elasticities.

The definition of  $R_i$  in equation (5) shows how unlikely it is that the distributional characteristics will be irrelevant. The ratios  $R_1$  and  $R_2$  will be equal only if 1) the marginal social utility of income is the same for all households, or 2) the relative quantities purchased of the two goods is the same for all households, or 3) some extremely improbable balancing of differences in quantities and social utilities occurs. In general, therefore, the relative optimal prices will reflect  $R_1$ ,  $R_2$ , and  $\lambda$ . Since  $\lambda$  is the shadow price of the budget constraint ( $\partial Z/\partial B = -\lambda$ ), the relative optimal prices will depend on the size of the deficit or surplus that the enterprise is required to have. This is in contrast to the Ramsey rule (equation (8)) in which the relative prices do not change with variations in the budget constraint.

In the special case of zero cross-elasticity of demand ( $\epsilon_{12} = 0$ ), it is easy to provide an intuitive interpretation of the role of the distributional characteristics and the budget constraint. Equation (7) now implies:

$$(9) \quad \frac{(p_1 - m_1)/p_1}{(p_2 - m_2)/p_2} = \frac{\epsilon_{22}}{\epsilon_{11}} \cdot \frac{(R_1 - \lambda)}{(R_2 - \lambda)}$$

This ratio of optimal "tax rates" or "rela-

tive profits" is the product of 1) an efficiency factor (the Ramsey ratio of price elasticities), and 2) a distributional equity factor. Since  $\epsilon_{22}/\epsilon_{11}$  is positive, the relative tax rates or profit rates vary with the corresponding  $R_i$ 's. The higher the value of  $R_i$ , i.e., the more that the consumption of the good is concentrated in low income families, the lower should be the relative price of that good. Equation (9) (and more generally equation (8)) provides a precise statement of how the Ramsey-Boiteux efficiency prices should be modified to reflect this principle of distributional equity.

The derivative of the optimal tax ratio of equation (9) with respect to  $\lambda$

$$(10) \quad \frac{\partial}{\partial \lambda} \left[ \frac{(p_1 - m_1)/p_1}{(p_2 - m_2)/p_2} \right] = \frac{\epsilon_{22}}{\epsilon_{11}} \frac{(R_1 - R_2)}{(R_2 - \lambda)^2}$$

shows that an increase (decrease) in  $\lambda$  raises (lowers) the relative price of good 1 if  $R_1$  exceeds  $R_2$ ; i.e., if the consumption of good 1 is more concentrated in low income families than the consumption of good 2. Since  $\lambda$  is the shadow price of the budget constraint, an increase in the required surplus raises  $\lambda$ . Therefore as the required surplus increases (or as the subsidy decreases), the price of the good with the lower income elasticity rises relative to the price of the good with the higher income elasticity.<sup>8</sup> Lower income families contribute an increased share of total revenue as a larger surplus is required.

## II. An Operational Measure of the Distributional Characteristic

This section suggests how an explicit operational expression for the distributional characteristic can be derived by adopting reasonable parametric forms for the three functional relations in terms of which  $R_i$  is defined:

<sup>8</sup> This change in relative prices never implies that the less expensive good becomes the more expensive unless the lower  $R_i$  becomes the higher  $R_i$ .

$$\begin{aligned}
 (11) \quad R_i &= \frac{N \int_0^\infty q_i(y) u'(y) f(y) dy}{Q_i \int_0^\infty q_i(y) u'(y) f(y) dy} \\
 &= \frac{N \int_0^\infty q_i(y) u'(y) f(y) dy}{N \int_0^\infty q_i(y) f(y) dy}
 \end{aligned}$$

First, let the demand relations have constant income elasticity of demand:

$$(12) \quad q_i(y) = b_i y^{\alpha_i}$$

A constant elasticity marginal social utility function,

$$(13) \quad u'(y) = y^{-\eta},$$

provides a convenient one-parameter representation of the normative welfare judgement.<sup>9</sup> The greater the value of  $\eta$ , the more egalitarian the implied social welfare function. Moreover, since this form implies that a 1 percent increase in income is associated with an  $\eta$  percent decrease in marginal social utility, the value of  $\eta$  has an intuitively natural interpretation.

These two assumptions imply that

$$(14) \quad R_i = \frac{\int_0^\infty y^{\alpha_i - \eta} f(y) dy}{\int_0^\infty y^{\alpha_i} f(y) dy}$$

or, by a change of variable from income ( $y$ ) to  $\log y$

$$(15) \quad R_i = \frac{\int_{-\infty}^\infty e^{(\alpha_i - \eta) \log y} g(\log y) d \log y}{\int_{-\infty}^\infty e^{\alpha_i \log y} g(\log y) d \log y}$$

The two integrals are equivalent to moment generating functions for the vari-

able  $\log y$  with "dummy" parameters  $\alpha_i - \eta$  and  $\alpha_i$ .

The distribution of income can be approximated quite well by the lognormal distribution. This approximation is particularly convenient in the current context. It implies

$$\begin{aligned}
 (16) \quad R_i &= \frac{e^{(\alpha_i - \eta) \bar{Y} + 1/2 (\alpha_i - \eta)^2 \sigma_Y^2}}{e^{\alpha_i \bar{Y} + 1/2 \alpha_i^2 \sigma_Y^2}} \\
 &= e^{-\eta \bar{Y} + 1/2 (\eta^2 - 2\alpha_i \eta) \sigma_Y^2}
 \end{aligned}$$

where  $\bar{Y}$  and  $\sigma_Y^2$  are the mean and variance of  $\log y$ . The lognormal distribution also implies a unique relation between the first two moments of  $\log y$  and the first two moments of  $y$ . The distributional characteristic can therefore be expressed as

$$(17) \quad R_i = \{\bar{y}^{-\eta} (1 + V)^{(1/2)\eta(1+\eta)}\} (1 + V)^{-\eta\alpha_i}$$

where  $\bar{y}$  is the mean of income and  $V$  is the relative variance,  $\sigma_y^2/\bar{y}^2$ .

Equation (17) shows that the distributional characteristic for each good can be written as the product of two terms, one of which depends on the parameters of the income distribution but not of the demand for the good. This general term

$$[\bar{y}^{-\eta} (1 + V)^{(1/2)\eta(1+\eta)}]$$

is a decreasing function of the average level of income and an increasing function of its relative inequality. It is in fact equal to the utility value of a dollar distributed uniformly among all the households, i.e.,

$$\bar{y}^{-\eta} (1 + V)^{(1/2)\eta(1+\eta)} = \int u'(y) f(y) dy$$

Of greater interest in the current context is the term that is specific to each good,  $(1 + V)^{-\eta\alpha_i}$ . This shows that a good with a higher income elasticity of demand ( $\alpha_i$ ) has a relatively lower value of  $R_i$ . A greater inequality of income in the population ( $V$ ) or a more egalitarian welfare function (i.e., a higher value of  $\eta$ ) reduces

<sup>9</sup> This isoelastic marginal utility function implies that the utility function is also isoelastic after the removal of an arbitrary constant. This form has been most common in the study of optimal growth.

$R_i$  for any value of  $\alpha_i$  and causes the  $R_i$  to be more sensitive to the distributional characteristics.

If the approximations suggested in this section are considered satisfactory for a particular problem, equation (17) provides a simple method of calculating each  $R_i$  in terms of available income distribution parameters ( $\bar{y}$  and  $\sigma_y^2$ ), an easily estimable income elasticity of demand ( $\alpha_i$ ) and an intuitively natural representation of the normative distributional judgement ( $\eta$ ). The optimal prices can then be calculated by solving equations (6a), (6b), and the budget constraint of equation (3).<sup>10</sup>

### III. Conclusion

The brief discussion in this paper has indicated how considerations of distributional equity can be included explicitly and operationally in the derivation of the optimal prices for public enterprises or regulated industries. A number of questions are left unanswered. For example, how should goods be priced that are sold to industrial firms rather than to households? What pricing rules are appropriate if price discrimination and multipart tariffs are permitted?<sup>11</sup> And how should the optimal pricing rule be modified if the demand is shifted to goods that are taxed or produced by public enterprises? The

growth of the public sector and the increasing concern with distributional equity make it important to develop a more complete theory of public pricing that incorporates considerations of distributional equity.

### REFERENCES

- A. P. Barten, "Consumer Demand Functions Under Conditions of Almost Additive Preferences," *Econometrica*, Jan. 1964, 32, 1-38.
- W. J. Baumol and D. F. Bradford, "Optimal Departures from Marginal Cost Pricing," *Amer. Econ. Rev.*, June 1970, 60, 265-83.
- M. Boiteux, "Sur la gestion des Monopoles Publics astreints a l'equilibre budgetaire," *Econometrica*, Jan. 1956, 24, 22-40.
- P. A. Diamond and J. A. Mirrlees, "Optimal Taxation and Public Production: II Tax Rules," *Amer. Econ. Rev.*, June 1971, 61, 261-78.
- A. K. Dixit, "On the Optimum Structure of Commodity Taxes," *Amer. Econ. Rev.*, June 1970, 60, 295-301.
- M. S. Feldstein, "Equity and Efficiency in Public Sector Pricing: The Optimal Two-Part Tariff," *Quart. J. Econ.*, forthcoming.
- , "The Pricing of Public Intermediate Goods," *J. Publ. Econ.*, forthcoming.
- A. C. Harberger, "Taxation, Resource Allocation and Welfare," in J. Due, ed., *The Role of Direct and Indirect Taxes in the Federal Revenue System*, Princeton 1964.
- J. R. Hicks, *Value and Capital*, 2d ed., New York 1946.
- A. P. Lerner, "On Optimal Taxes with an Untaxable Sector," *Amer. Econ. Rev.*, June 1970, 60, 284-94.
- H. Mohring, "The Peak Load Problem with Increasing Returns and Pricing Constraints," *Amer. Econ. Rev.*, Sept. 1970, 60, 693-705.
- F. Ramsey, "A Contribution to the Theory of Taxation," *Econ. J.*, Mar. 1927, 37, 47-61.

<sup>10</sup> The relevant price elasticities of equation (6) cannot remain constant if the basic Slutsky relation is to be satisfied. Equation (6) is correct for the elasticity values that prevail at the optimum point. In practice, constant values may be an adequate approximation within the relevant range. If not, some more general demand structure must be considered; see, e.g., A.P. Barten.

<sup>11</sup> See Feldstein for the theory of the optimal two-part tariff and a specific application.

# The Industrial Composition of U.S. Exports and Subsidiary Sales to the Canadian Market

By THOMAS HORST\*

This paper explores the empirical relationship between *U.S.* exports and foreign direct investments. Since the first section of this paper bears a resemblance to the study by William Gruber, Dileep Mehta, and Raymond Vernon, let me begin by comparing our two approaches. Exporting and foreign investing, according to their analysis, are separate stages in the dynamic process by which *U.S.* firms expand into foreign markets. A firm competing in the large and technically sophisticated *U.S.* market is induced to develop a product superior to that of a firm selling to a smaller and less sophisticated market in a foreign country. At the early stages of the product's development, the *U.S.* firm can exploit this technological advantage over its foreign competition by exporting to the foreign market. But as time moves on, the foreign market expands, and the advantages to *U.S.* production diminish. The *U.S.* firm, wishing to maintain its initial share of the foreign market, is eventually forced to establish marketing, service, and production facilities in the foreign country. Foreign investment is thus the successor to foreign trade.

This paper differs from Gruber, Mehta, and Vernon in its emphasis on the static,

rather than the dynamic, relationship between *U.S.* exports and foreign investment. My objective is to show how static considerations such as technological knowledge, tariff rates, market size, and factor costs influence the export and direct investment decisions of *U.S.* firms. Since the analyses of the static and dynamic aspects of an issue are often complementary, there need be no fundamental conflict between my findings and those of Gruber, Mehta, and Vernon. But to the extent that my analysis may convince the reader that static elements do influence trade and investment decisions, the notion of a *rigid* cycle of foreign market exploitation—a sequence of events which can be neither hastened nor delayed, much less reversed—will be weakened.

My empirical analysis focuses primarily on the fairly recent (1963) situation of *U.S.* manufacturing firms selling to the Canadian market, largely because available data permit a fuller investigation of this situation than that for any other year or for any other country. My first significant finding is that the technological intensity of a *U.S.* manufacturing industry is more closely related to the *sum* of that industry's exports to Canada and its subsidiary sales in Canada than it is to either exports or subsidiary sales taken separately. Such a finding clearly supports the notion that exporting and foreign direct investing may be substitutes for one another and suggests, somewhat more generally, that studies of the commodity

\* Department of economics, Harvard University. This paper is derived from my Ph.D. dissertation written at the University of Rochester. I have benefited from considerable criticism along the way, especially from Ronald Jones, Rudolph Penner, Richard Caves, and two referees for this *Review*. Financial support from a Ford International Grant to Harvard University is gratefully acknowledged.

composition of *U.S.* foreign trade would do well to take account of foreign investment opportunities. My second significant finding shows that Canadian tariff policy has had a definite impact on the choice between exporting and Canadian subsidiary production—the higher the Canadian tariff, the smaller the share of *U.S.* exports and the larger the share of Canadian subsidiary production in total *U.S.* sales to the Canadian market. In this latter context I have employed estimates of both the nominal and the effective rates of Canadian tariff protection. Although the effective rate comes closer than the nominal to measuring the incentive to substitute subsidiary production for exports, the actual extent of the substitution is predicted equally by nominal and effective rates. When an analogous, though far more limited, experiment is performed using data for the United Kingdom and the Common Market, the nominal rate is superior to the effective rate in predicting the substitution of subsidiary production for exports. All of this calls into question the many assumptions made in calculating effective rates of tariff protection, if not the very notion of an effective rate itself.

In investigating any topic as complex as the industrial composition of *U.S.* exports and foreign direct investments, several analytical simplifications must be made. I have chosen to divide this study into two sections, the first concentrating on how much *U.S.* firms will be able to sell abroad either by exporting or by establishing a foreign subsidiary, the second exploring how *U.S.* firms choose between these two alternative forms of foreign market participation. Although in a fully specified model of foreign trade and investment behavior all variables would be determined simultaneously, the format I have chosen simplifies the analysis considerably and yet allows me to focus on the

substitutional nature of foreign trade and investment opportunities.

### I. Determinants of Total *U.S.* Sales to the Canadian Market

This section investigates the determinants of total *U.S.* sales to the Canadian market and leaves for the following section the analysis of how *U.S.* firms choose between exports and subsidiary production in supplying the Canadian market. The theoretical rationale for the regression equations described below derives from a specific notion of the role of technology in the production process. Technology is assumed to have the essential characteristics of a public good<sup>1</sup>—that is to say, once a technology has been developed by a firm, it can be applied by many plants in different locations at the same time. But unlike the usual public good, technology can be prevented by patent rights or by safeguards against industrial espionage from flowing to unauthorized users. While the firm that develops the technology might sell its knowledge to outsiders, it may also choose to retain that technology in the hope of capturing directly the full monopolistic rent to its know-how.

The following analysis is predicated on the assumption that *U.S.* manufacturing firms selling to the Canadian market in 1963 were willing and able to prevent their Canadian competitors from expropriating *U.S.* technological knowledge and that this technological advantage could be exploited either by producing in the United States for export to Canada or by transferring their technology to a Canadian subsidiary.<sup>2</sup> The principal purpose of this section is to show that the substitutional

<sup>1</sup> This role of technology is very clearly formulated by Harry Johnson.

<sup>2</sup> A. E. Safarian (1966, 1969) has presented convincing evidence that the technology of *U.S.* parent firms tends to be fully available to their Canadian subsidiaries.

relationship between exporting and investing abroad, stemming from the applicability of the same technology to production in either country, can in fact be observed in the cross-industry patterns of *U.S.* exports and subsidiary sales to the Canadian market.

Translating this notion of the substitutability of direct investment for exporting into a testable hypothesis is no easy task. The first challenge is to devise some measure by which a *U.S.* industry's success in selling to the Canadian market can be properly judged. There are several standards one could conceivably devise—I have chosen to compare a *U.S.* industry's exports and its Canadian subsidiary production to the sum of world exports to Canada plus total Canadian production for the industry in question. This choice is especially significant given both the abundance of Canadian natural resources and the Canadian dependence on manufactured imports from various foreign countries. By including total Canadian production in the standard for comparison, I am effectively taking the existence of Canadian resources as given and considering the more interesting question of why these resources may be exploited by *U.S.* rather than by Canadian or other foreign firms. By including world exports to Canada in the standard, I will be implicitly comparing *U.S.* sales not only to those of other firms producing in Canada, but also to those British, Japanese, and other foreign firms exporting to Canada. This broad measure, which I will call "the Canadian market," should give a fair impression of how *U.S.* firms do relative to all other firms selling in Canada.

A second and far more difficult challenge is to devise a satisfactory measure of the technological advantage of *U.S.* firms relative to their Canadian competitors. In theory the problem seems insurmount-

able; not only must we combine product technology with the more general organizational, marketing, and financial skills of *U.S.* management, but we must also be careful to distinguish the technology existing at a particular point in time (a stock) from the most recent contributions to that technology (a flow). In practice the problem may be less formidable: as one compares one industry to another, not only do the various bases of technological advantage appear to be multicollinear, but also their relative importance seems to be reasonably stable over time.<sup>3</sup> There is some justification, therefore, for my taking new company-sponsored research and development expenditures deflated by *U.S.* sales as a rough proxy for the more broadly based technological advantage *U.S.* firms enjoy over their competitors for the Canadian market.

How does technological intensity affect a *U.S.* industry's share of the Canadian market? In the following equations the dependent variables are the 1963 shares of *U.S.* exports,  $X_i$ , subsidiary production,  $P_i$ , and the sum of the two,  $S_i$ , in world exports to Canada plus total Canadian production for eighteen two-digit manufacturing industries. Each share was regressed on the 1963 company-sponsored R&D expenditures as a share of domestic sales in the United States with the following results:<sup>4</sup>

$$X_i = .043 + 5.47 R\&D_i \quad R^2 = .33 \text{ d.o.f.} = 16 \\ (2.69)$$

$$P_i = .166 + 14.65 R\&D_i \quad R^2 = .48 \text{ d.o.f.} = 16 \\ (3.71)$$

$$S_i = .209 + 20.69 R\&D_i \quad R^2 = .63 \text{ d.o.f.} = 16 \\ (5.10)$$

<sup>3</sup> Gruber, Mehta, and Vernon present some quantitative support of this assumption.

<sup>4</sup> Descriptions of the sources and adjustments of the data on which this study is based are contained in the Appendix.

The number in parentheses is the  $t$ -statistic of the above coefficient. That each of these regressions would yield a statistically significant fit comes as no surprise. What is interesting is how much better the fit of the third equation is than the fit of either of the previous two. That R&D is more closely related to *total U.S.* sales to the Canadian market than to either *U.S.* exports or production by *U.S.-owned* subsidiaries provides strong, if indirect, support for the hypothesis that exporting and foreign investing represent alternative methods by which *U.S.* firms exploit the same technological advantages over their Canadian competitors. An obvious corollary of this finding is that the commodity composition of *U.S.* exports would be better explained if some account were taken of the distribution of *U.S.* direct foreign investments. To the best of my knowledge no study of the commodity composition of *U.S.* foreign trade has attempted to deal with this factor explicitly.

## II. Determinants of the Choice Between *U.S.* Exports and Foreign Subsidiary Production

Having shown that exporting and subsidiary production may be alternative methods of exploiting the same technological advantages in selling abroad, we are still left with the important question of how *U.S.* firms choose between these two methods of supply. In theory, the choice is simple: for any given volume of foreign sales, a firm should try to minimize its aggregate supply costs. In practice, this means that we should ignore the technological advantages applicable to both exporting and subsidiary production and concentrate instead on the barriers to trade and the differentials between the two countries in the real costs of production. We are left, in short, with a relatively straightforward application of the tradi-

tional comparative-cost theory of international trade analysis.

Notice that we are working with two very different notions of comparative advantage in this paper. The first and more recent notion is the technological advantage enjoyed by *U.S.-owned* firms (be they in the United States or be they *U.S.-owned* subsidiaries in foreign countries) over their foreign-owned competitors; the second and more traditional notion is the real cost advantages enjoyed by plants *located in the United States* over plants located within foreign countries. The former notion is crucial in the first, but not the second, part of this investigation, while the latter notion is crucial in the second, but not the first, part of the analysis.

In the particular instance of *U.S.* firms selling to the Canadian market in 1963, the principal factors determining the cost of exporting relative to the cost of Canadian subsidiary production would seem to be:

- 1) the Canadian tariff on the exports, but not the subsidiary production, of *U.S.* firms;
- 2) the higher cost of manufactured inputs in Canada compared to their cost in the United States;
- 3) the lower cost of many natural-resource inputs in Canada compared to their cost in the United States;
- 4) the lower wage paid to Canadian workers compared to that paid to *U.S.* workers; and
- 5) the higher costs of small-scale production characteristic of Canadian but not of *U.S.* plants.

Quantifying the importance of these cost considerations is a difficult task. James Melvin and Bruce Wilkinson have studied the Canadian tariff structure, and from their findings I have constructed an estimate of the nominal tariff applied to the

import of finished goods for each industry.<sup>5</sup> If the higher cost of Canadian material inputs is a reflection of the tariffs on their import, we can also make good use of Melvin and Wilkinson's estimates of the effective rates of protection. An effective rate of protection, one may recall, measures the net protection given to a domestic industry by the whole tariff system. Since net protection is the difference between the tariff on finished imports and the tariff-induced increase in the cost of necessary material inputs, an effective rate of protection should reflect the combined impact of cost considerations 1) and 2) above on the export vs. subsidiary production choice for a U.S. firm. The regression equations presented below show the results of using first the nominal and then the effective rate of tariff protection to predict the form of a U.S. industry's participation in the Canadian market.

The impact of the lower cost of natural resource and labor inputs in Canada as well as the diseconomies of small-scale Canadian production are considerably more difficult to estimate. One should not, however, succumb to the temptation to ignore these real-cost considerations. The tendency of tariff setters to give more protection wherever more protection is necessary to assure the survival of a domestic industry would indicate that real-cost differentials between Canada and the United States may be positively correlated with the tariff encouraging Canadian production. Ignoring real-cost considerations would thus lead to an *underestimate* of a tariff's ability to discourage exporting and encourage subsidiary production. What I have done is to construct a variable,  $Z_i$ , defined to be the ratio of Canadian market size (Canadian production plus imports) to

U.S. domestic production for industry  $i$ . The purpose of including Canadian imports in the numerator of  $Z_i$  is to capture insofar as possible the potential economies of scale which a manufacturer would achieve in substituting subsidiary production for his current exports. Assuming that  $Z_i$  is relatively large for those industries in which the real-cost advantages of Canadian production are relatively large, including  $Z_i$  in the regression equations should lead to a more honest estimate of the significance of the tariff factor. Notice that including Canadian imports in  $Z_i$  will *not* introduce any spurious correlation into the regression results, since  $Z_i$  is presumed to have a *negative* impact on the share of U.S. exports in total U.S. sales to the Canadian market.

The dependent variable (before the logarithmic transformation) in the following regression equations is the share of U.S. exports in total U.S. sales to the Canadian market. The variables  $t_i$  and  $e_i$  are the estimated nominal and effective rates of tariff protection, respectively.

$$\ln\left(\frac{X_i}{S_i}\right) = -.23 - \frac{2.45}{(3.23)} \ln(1+t_i) - \frac{12.6}{(3.24)} Z_i$$

$$R^2 = .70$$

$$\text{d.o.f.} = 15$$

$$\ln\left(\frac{X_i}{S_i}\right) = -.38 - \frac{1.20}{(3.44)} \ln(1+e_i) - \frac{11.3}{(2.90)} Z_i$$

$$R^2 = .71$$

$$\text{d.o.f.} = 15$$

All coefficients have the expected negative sign and are significantly less than zero at the 1 percent level. These results are sufficiently novel, I believe, to merit both a justification of the logarithmic transformation and a comparison of the two equations based on different concepts of tariff protection.

The logarithmic transformation was not

<sup>5</sup> I am deeply indebted to Melvin and Wilkinson for furnishing me with the Canadian import data necessary to aggregate their tariff estimates for my purposes (see Appendix for details).

specified a priori, but only after the tendency of export shares to respond much more sharply to an increase in a low tariff than to an increase in a high tariff was observed. This non-linear relationship could be due to a non-linear distribution of the underlying cost functions or the tendency of high tariff rates to give redundant protection. This nonlinearity may also be due to the way a profit-maximizing firm determines the optimal transfer price for its exports. If the Canadian tariff is quite high, a *U.S.* firm may wish to channel its exports through a wholly owned Canadian distributor in order to minimize the transfer price and avoid the full burden of the Canadian tariff. But if the Canadian tariff is low and the *U.S.* firm is anxious to show a profit in the United States rather than in Canada, exports may be declared at their full market value and sold directly to the Canadian consumer. Thus, even if the relationship between Canadian tariffs and the volume of Canadian imports were linear, the measured relationship between the tariff and the *value* of these imports could be non-linear.

Judging from the  $R^2$ 's of the two preceding equations, the nominal rate of protection serves as well as the effective rate in predicting the extent to which subsidiary production will be substituted for exports. Since the effective rate was intended to yield a better estimate of the protection given to domestic production by a country's tariff system, this tie score must be disappointing to those who would go to the considerable trouble of calculating effective rates of protection. Although this result could be spurious, I suspect it is not. I am skeptical in particular of the assumptions made in calculating effective rates—to wit, the assumptions that all production is subject to constant returns to scale and that domestic input costs will equal world prices times the nominal tariff factor. Neither of these assumptions are in close

accord with the available evidence. Several empirical studies<sup>6</sup> have concluded that Canadian production tends to be less efficient than *U.S.* production, the inefficiency usually being attributed to the small scale of most Canadian production. If so, a large foreign investor need not take Canadian input costs as given, since his own demand may stimulate low-cost production of necessary inputs.

The assumption that Canadian prices are bid up to the full extent of the Canadian tariff has not been fully tested, but here too the available evidence suggests that Canadian prices may be somewhat lower than what the existing tariff protection permits.<sup>7</sup> Although very little is known about the transfer pricing policies of multinational firms, the possibility that *U.S.* investors may be exporting parts and materials to their Canadian subsidiaries at prices below the world price serves to undermine further this crucial assumption used in calculating effective rates of protection. Unfortunately, it is difficult, if not impossible, to gather the information one would need to correct the estimates of the effective rates of protection for these known or suspected deficiencies.

The higher coefficient of the nominal rate of protection relative to that of the effective rate is easy to account for statistically but somewhat more difficult to assess economically. There is a very high correlation between nominal and effective rates of protection (.83 in the Melvin and Wilkinson study), and the standard deviation of the effective rates is more than double that of the nominal rates. When the two variables are used in identical regression equations, one naturally finds that the effective tariff coefficient is roughly half the size of the nominal tariff coefficient. But this does not answer the economically

<sup>6</sup> See Harry English, Harry Eastman and Stefan Stykolt, and Ronald Wonnacott and Paul Wonnacott.

<sup>7</sup> See Wonnacott and Wonnacott, ch. 14.

relevant question of how large a substitution of subsidiary production for exports would be achieved by raising any given tariff rate by 1 percent. Do nominal rates underestimate the significance of Canadian tariff protection, or do effective rates overestimate its significance? Until more accurate estimates of the effective rate of protection are obtained, there can be no very satisfactory answers to these important questions.

The principal conclusion suggested by these regression results is that Canadian tariffs have influenced the willingness of *U.S.* firms to produce within Canada rather than exporting to Canada. To see if this conclusion holds elsewhere, I have run analogous regressions for *U.S.* exports and foreign investments in the United Kingdom and Common Market. The chief difference between these results and the preceding results for Canada is that a complete set of data is available for only seven industries (paper, rubber, metal products, machinery, electrical machinery, transportation equipment, and chemical products) and that no data could be gathered to construct the  $Z_i$  variable. Regressing the share of *U.S.* exports in total *U.S.* sales on estimates of the nominal and effective rates of protection taken from Bela Balassa's study yields the following equations for the United Kingdom:

$$\frac{X_i}{S_i} = 2.66 - \frac{2.14}{(4.37)} (1 + t_i) \quad R^2 = .79 \text{ d.o.f.} = 5$$

$$\frac{X_i}{S_i} = 1.27 - \frac{0.85}{(2.93)} (1 + e_i) \quad R^2 = .63 \text{ d.o.f.} = 5$$

For the Common Market I obtain the following:

$$\frac{X_i}{S_i} = 4.40 - \frac{3.75}{(2.62)} (1 + t_i) \quad R^2 = .58 \text{ d.o.f.} = 5$$

$$\frac{X_i}{S_i} = 1.06 - \frac{0.65}{(0.94)} (1 + e_i) \quad R^2 = .15 \text{ d.o.f.} = 5$$

Notice that in these equations the performance of the effective rates of protection is actually worse than that of the nominal rates of protection. This suggests the unpleasant possibility that calculating effective rates of protection may in some instances be counterproductive. If large internal markets serve to undermine the assumption that domestic input costs equal world prices times the nominal tariff factor, then one would expect effective rates perform best in Canada, somewhat worse in the United Kingdom, and worst of all in the Common Market—which is the ranking of their success in this study.

### III. Conclusion

The principal findings of this study are that exporting and subsidiary production represent alternative means by which *U.S.* firms exploit technological advantages over their foreign competitors and that tariffs imposed by the foreign country encourage *U.S.* firms to substitute subsidiary production for exporting. In this latter context I have experimented with estimates of both the effective and the nominal rates of tariff protection. Although in theory an effective rate of protection should give a better estimate of the tariff-induced barrier to exporting, in practice the nominal rate predicts the actual substitution of subsidiary production for exporting at least as well. The most plausible explanation for this discrepancy between theory and practice seems to be that foreign production is not subject to constant returns to scale and the foreign cost of material inputs does not accurately reflect the tariff on their import, facts which contradict the assumptions made in calculating effective rates of protection.

In focusing on these particular aspects of the foreign trade and investment process, I have necessarily ignored many important issues. Among these deliberate omissions are various considerations sug-

gested by oligopolistic behavior and general equilibrium analysis,<sup>8</sup> considerations which a researcher with better data or different objectives would surely want to investigate. But while this study has been limited to a narrow set of issues, it does lead to one general conclusion. When large corporations become a country's principal exporters and predominant foreign investors, perfect competition is no longer a good working assumption for the empirical analysis of international economic phenomena. Taking the elements of monopolistic and oligopolistic behavior into explicit account may lead to new insights into the foreign trade and investment process.

#### STATISTICAL APPENDIX

Described below are the sources and transformations of the data on which the preceding analysis is based. Since the Canadian and Western European data come from separate sources, let me begin with the Canadian data. All variables are derived from annual data for 1963, except for the tariff estimates which were based on Canadian import data for 1962:

$P_i$  ≡ Sales by U.S.-owned subsidiaries in Canada divided by total sales of all Canadian firms plus total Canadian imports for commodity  $i$ . Sales data are taken from the *Corporations and Labour Unions Returns Act, Report for 1963* and converted to U.S. dollars at the current exchange rate, 0.927 US\$/Can\$. Canadian import data are taken from the O.E.C.D., *Foreign Trade Statistical Bulletins, Series B*. Jan.-Dec. 1963.

$X_i$  ≡ U.S. exports to Canada divided by total Canadian sales plus total Canadian imports of commodity  $i$ . Data sources are the same as those for  $P_i$ .

$S_i$  ≡  $X_i + P_i$

$R\&D_i$  ≡ Company-sponsored research and development expenditures divided by total sales for industry  $i$  in the United States. R&D data were taken from the National Science Foundation, *Research and Development in Industry, Annual Report for 1963*. American sales data are from the Department of Commerce, *1963 U.S. Census of Manufactures*. In six industries (beverage, tobacco, leather, wood, furniture and fixtures, and printing and publishing) R&D expenditures were not reported separately. The estimate of  $R\&D_i$  for the beverage industry is based on the 1962 value for the food and beverage industries. The remaining five industries were given the rate, 0.0010, of all "other industries." Since the total expenditure on R&D by these industries was very small relative to their total sales, I doubt that any significant distortions were introduced with this approximation.

$t_i$  ≡ Estimated nominal protection of the Canadian industry. This estimate is based on the rates calculated by Melvin and Wilkinson. Their study contains an estimate of the nominal rate and two estimates of the effective rate of protection for 133 three-digit manufacturing industries in Canada. Since my study is based on 17 two-digit industries, I had to average the rates of the component industries. The authors have very generously furnished me with the import data necessary to take the appropriate weighted averages.

$e_i$  ≡ Estimated effective rate of protection in Canada. This estimate was the import-weighted average of the Melvin and Wilkinson estimates using the 5 percent imputed tariff on unspecified inputs.

<sup>8</sup> Richard Caves has put many of these considerations into the coherent framework of industrial organization analysis.

The Western European data for sales by U.S.-owned subsidiaries, unlike the Cana-

dian data, were based on the U.S. Department of Commerce, *Survey of Current Business*, European tariff estimates were based on Bela Balassa's study and were converted to the Standard Industrial Classification according to the following scheme:

TABLE 1—ASSUMED CORRESPONDENCE BETWEEN  
BALASSA'S CLASSIFICATIONS AND THE S.I.T.C.

Industry	Balassa No.	Assumed S.I.T.C. No.
Paper	32	64
Chemicals	38	58
	42	55
	40	5-55-58
Rubber	37	62
Metals and	48	671
Metal Products	49	672
	50	673+674
	55	679
	51	675+...+678
	54	68
	56	69
Machinery	57	712
	58	71-712
Electrical		
Machinery	59	72
Transportation	60	735
Equipment	61	731
	62	732
	64	733
	65	734

Note: Following Balassa, I weighted the averages of these tariffs by the exports of the ten industrial countries covered in Balassa's study.

#### REFERENCES

- B. Balassa, "Tariff Protection in Industrial Countries: An Evaluation," *J. Polit. Econ.*, Dec. 1965, 73, 573-94.
- R. E. Caves, "International Corporations: The Industrial Economics of Foreign Investment," *Economica*, Feb. 1971, 38, 18-44.
- H. C. Eastman and S. Stykolt, *The Tariff and Competition in Canada*, Toronto 1967.
- H. E. English, *Industrial Structure in Canada's International Competitive Position*, Montreal 1964.
- W. Gruber, D. Mehta, and R. Vernon, "The R&D Factor in International Trade and International Investment of U.S. Industries," *J. Polit. Econ.*, Feb. 1967, 75, 20-37.
- T. Horst, "A Theoretical and Empirical Analysis of American Exports and Direct Investments," unpublished doctoral dissertation, Univ. Rochester 1969.
- , "The Theory of the Multi-national Firm: Optimal Behavior under Different Tariff and Tax Rates," *J. Polit. Econ.*, Sept. 1971, 79, 1059-72.
- H. Johnson, "The Efficiency and Welfare Implications of the International Corporation," in C. P. Kindleberger, ed., *The International Corporation*, Cambridge 1970.
- J. Melvin and B. Wilkinson, "Effective Protection in the Canadian Economy," Special Study No. 9, Economic Council of Canada, Ottawa 1968.
- A. Safarian, *Foreign Ownership of Canadian Industry*, Toronto 1966.
- , *The Performance of Foreign-owned Firms in Canada*, Montreal 1969.
- R. Wonnacott and P. Wonnacott, *Free Trade Between the United States and Canada*, Cambridge 1967.
- Dominion Bureau of Statistics, *Corporations and Labour Unions Returns Act, Report for 1963*, Ottawa 1967.
- Organization for Economic Cooperation and Development, *Foreign Trade Statistical Bulletins*, Series B, Jan.-Dec. 1964.
- U.S. Department of Commerce, 1963 *U.S. Census of Manufacturers*, Washington 1966.
- , *Survey of Current Business*, Nov. 1966, 46, 7-11.
- U.S. National Science Foundation, Basis Research, *Applied Research and Development in Industry*, 1963, Washington 1966.

# The Process Analysis Alternative to Statistical Cost Functions: An Application to Petroleum Refining

By JAMES M. GRIFFIN\*

Contrary to theoretical expectations, most statistical cost function studies show that "marginal cost is constant."<sup>1</sup> On the basis of results for the *U.S.* petroleum refining industry, this paper questions the above conclusion and suggests an alternative approach. Specifically, a process analysis approach,<sup>2</sup> rather than a statistical cost function approach, yields the classical short-run cost function properties—rising marginal costs and a U-shaped average cost function. Also, this study demonstrates the generality of the process analysis approach by first treating the single and then the joint product<sup>3</sup> cases. The application to joint production is particularly promising because many industries involve aspects of joint production and yet the treatment of joint product cost functions in the literature is an area of little progress.<sup>4</sup>

\* Economist, Standard Oil Company (New Jersey). Much of this paper is drawn from my dissertation submitted to the University of Pennsylvania, May 1971. I gratefully acknowledge indebtedness to F. Gerard Adams, L.R. Klein, and Almarin Phillips for numerous helpful comments.

<sup>1</sup> See A. A. Walters, p. 46, and John Johnston.

<sup>2</sup> While Alan Manne does not consider the cost function implications of process analysis, his pioneering study serves as a basis for the present extension of process analysis.

<sup>3</sup> By joint production, I mean the more contemporary definition of products produced with interdependent production techniques and a variable product mix.

<sup>4</sup> See Johnston's discussion of this problem, p. 185. Also for the related problem of estimation of joint production functions see Hrisikesh Vinod, Phoebus Dhrymes and Bridger Mitchell.

George Green's estimates of the short-run average cost function for petroleum refining establishments are typical of most statistical cost studies. Green estimated average cost functions for *U.S.* petroleum refining establishments of various sizes using census cross-section data from 1954–61. The study incorporates variables to reflect changes in the product mix as well as an aggregate capital variable to correct for the "regression fallacy."

With regard to the short-run cost curve, Green concluded: "All plants, regardless of size, show an L or \ shaped pattern of average cost over the observed ranges of output for equations relating to 1957 and later years" (p. 122).

This result also applies for large plants prior to 1957. Only small plants reflect a mild U-shaped pattern for 1954–56. He attributes the declining or constant average costs either to under-utilization of capacity during the sample period or to production processes, which give a short-run average cost function with an L shape rather than a U shape.

In analyzing similar empirical findings, Johnston lists several criticisms of statistical cost analysis. These criticisms include the possibility of observations over a narrow output range, simultaneous equation bias, deflation of inputs and outputs, etc. Any of these can affect the shape of the estimated cost function. However, as Johnston points out, the importance of

these factors is not known nor in most cases is it clear that these factors bias the result towards a flat cost function. Because these criticisms remain on a theoretical level, essentially untested by empirical observation, they lose much of their force. Perhaps the most disturbing fact is that the empirical results of statistical cost functions are generally incompatible with the theory of perfect competition.

The purpose of this paper is to investigate whether the process analysis approach to cost curves offers new insights into cost curve analysis and yields results which are more in accord with theory. It should be emphasized that this is a short-run analysis since the configurations of capital equipment are fixed; therefore, questions involving long-run cost functions are not considered.

Methodologically, the statistical cost function and process analysis approaches are quite different. While the statistical cost function approach focuses directly on sample observations of costs and outputs to estimate the cost function, process analysis is indirect in that the emphasis is placed on describing the production function from engineering data. After describing the production function, the cost-output relationship is then derived as a result of assumed optimization behavior. For example, in the single product case, the cost function can usually be derived by minimizing the cost of various output levels given factor prices and the production function. In process or activity analysis, the production function is described in a detailed linear programming (*LP*) manner where each of the column vectors describes a processing activity in engineering or physical terms. One of the disadvantages of an *LP* description of the production function, which the statistical cost function approach does not face, is that the matrix of technical coefficients is fixed for all output levels. However, this

problem can be overcome by approximating non-linear functions by linear segments.<sup>5</sup> The activity vectors attempt to describe the possible options open to the refiner in selecting the optimal production combinations. Equation constraints include material balance equations, the limitations of raw material inputs, the quality specifications on the products produced, and the available capital, which is disaggregated into process types.

It is useful to ask why the process analysis approach which is over 20 years old tends to be overlooked by most econometricians. Do the reasons relate to the methodological difference between using engineering data and accounting data? Are the former viewed as subjective while the latter are viewed as objective? Such a distinction loses much force for the typical econometric analysis using proxy variables and a functional form selected on the basis of fit. Besides the type of data, the quantity and relative scarcity of engineering data compared to accounting data may also explain the neglect of the process analysis approach. Also process analysis requires a greater technical knowledge of the industry and greater computational effort than the statistical cost function approach.

A discussion of the application of process analysis to the single and joint product cases serves both to illustrate its flexibility and to explain the factors producing results contrary to the evidence found in statistical cost function studies. Before considering the single and joint product cases, relevant details about the production function, the aggregation conditions, and the treatment of technological change are introduced in Section I. In Section II, a process analysis approach is used to derive

<sup>5</sup> For example, in petroleum refining the addition of lead reacts nonlinearly with octane. The curve can be approximated by building separate activity vectors to account for the effect of lead on octane for various levels of lead; e.g., 0-5cc, .5-1.0cc, 1.0-2.0cc, etc.

the cost curve in the single product case. Section III extends the analysis to the joint product cost curve emphasizing the dependency of product outputs on relative product prices. Section IV briefly summarizes the major conclusions.

### I. The Inputs for Process Analysis

Refiners use *LP* models to describe their production function for their monthly refinery scheduling. This particular application will focus on the industry level rather than the firm level; nevertheless, the approach remains essentially similar. The industry production function can be written as a weighted average of *LP* models for each refinery in the United States as follows:

$$(1) Ax \leq b$$

where:

$$Ax = \sum_{i=1}^j A_i x_i \text{ and } b = \sum_{i=1}^j b_i$$

and the  $i$ th subscript refers to  $i$ th refinery.

The  $n \times 1$  vector,  $x$ , gives the possible production activities. The  $A$  matrix of technical coefficients is of dimension  $m \times n$  in which the  $a_{ij}$ th element gives the amount of the  $i$ th constrained resource consumed (produced) per unit of activity  $x_j$ . The term  $b$  is an  $m \times 1$  vector of constrained elements specifying the various capital process availabilities, product quality specifications, and material balance constraints. Unfortunately, the  $A_i$  and  $x_i$  for each refinery are not known, nor would it be economically feasible to obtain such data; therefore, the Marshallian concept of the representative firm is particularly useful in describing the industry production function.

The *LP* chosen here to represent the industry is a modified Gulf Coast 200,000 barrel per day refinery *LP* developed by Bonner & Moore Associates for the American Petroleum Institute's study of the

costs of reducing the air pollution qualities of gasoline.<sup>6</sup> To make this *LP* model representative of the industry, modifications include an average U.S. crude type plus blending options and process equipment found in other refineries of different size and location. The detail of the *LP* (227 equations and 342 processing options) partially reflects the effort to make it representative of the industry.<sup>7</sup>

In measuring cost curves over time, it is necessary to allow explicitly for the effect of technological change as reflected in new process and product improvements. Changes in product spectrum and product quality specifications can be updated continuously. New process techniques resulting from one process replacing another are stated explicitly by changing the  $b$  vector to show the various quantities of process types available for production. Unfor-

<sup>6</sup> The Bonner & Moore model is based on engineering data on the actual operation of the various processes under possible operating conditions. For example, the catalytic cracking process can operate with different feedstocks, temperatures, recycle ratios, and catalyst conditions. Similarly, the blending equations are set up such that any combination of blending stock can be used as long as the product quality specifications are met. The quality of the data is good because of Bonner & Moore's close affiliation with a major builder of refinery units.

<sup>7</sup> The industry level cost function derivation involves several complications avoided at the individual refinery level. The use of a representative refinery to describe the industry production function involves aggregation over space and production techniques. First, the aggregation over production techniques requires that the industry's weighted average matrix of technical coefficients is approximated by the technology matrix of the representative plant ( $A_i$ ).

Second, spatial aggregation over the factor inputs is implied in the use of an aggregate industry production function. In summing the process capital of each type all refineries in the  $b$  vector, one assumes, in effect, that while capital processes are physically separated, capital services are transferable between refineries. To a large degree, capital services are transferable since refiners trade finished and semifinished products through the maze of pipelines to alleviate imbalances in the capital processes. For the small refineries operating outside of the nation's eight large refining areas, such trading may be severely restricted by costs.

tunately, data are not available to distinguish improvements among the various models within a process type.

## II. The Single Product Case

As an example of a single product case, it was assumed that the U.S. petroleum refining industry produces a basic set of products in fixed proportions. The process analysis statement of the problem is to minimize total variable costs for various output levels ( $b_j^i$   $i=1 \dots p$ ) along the fixed product mix as follows:

$$(2) \quad \text{minimize } c'x^i \quad \text{for } i = 1 \dots p$$

$$\text{subject to } Ax^i \leq b^i$$

where:  $b^i = (b_1 \dots b_j \dots b_m)'$

and:  $b_j^i = b_j + i\delta$

$\delta$  = increment in output  
for each observation  
along the cost curve

The  $1 \times n$  vector of operating costs,  $c$ , shows for the  $c_j$ th element the cost per unit of using the  $x_j$ th activity. The  $x_k$ th activity denotes the output of the fixed product mix, while the  $a_k$  column vector of the  $A$  matrix specifies the product yields and the  $b_j$ th element of  $b$  determines the output of the fixed product mix.<sup>8</sup>

The solution to equation (2) requires a specification of  $c$ ,  $A$ , and  $b$ . A simple linear programming algorithm will give both the primal ( $x$ ) and dual ( $y$ ) solutions and the objective function, which is equivalent to total variable costs for each specific output level. By adding fixed costs ( $FC$ )<sup>9</sup> to total variable costs, one obtains the short-run total cost function. The short-run average cost function can then be calculated as follows:

<sup>8</sup> The  $b_j$ th constraint is rewritten to appear as  $a \geq$  constraint.

<sup>9</sup> Fixed costs include capital charges, maintenance, overhead, and labor costs. Capital charges are based on the 1965 construction costs for the various process units given by Bonner & Moore. For a description of this calculation, see Griffin, Appendix B.

$$(3) \quad SRAC = \frac{c'x^i + FC}{x_k^i} \quad i = 1 \dots p$$

Because the objective function for an optimal solution is equal for both the primal and dual solutions, the dual of the  $b_j$ th element may be interpreted as the short-run marginal cost of an additional output of the fixed product mix as follows:

$$(4) \quad \text{Total Variable Costs} = c'x^i = b'y^i$$

$$i = 1 \dots p$$

$$(5) \quad SRMC = \frac{d \text{ TVC}}{d b_j^i} = y_j^i$$

As inputs, the LP requires that all fixed factor inputs such as the engineering capacities for each of the twelve processes, which are reported by the *Oil and Gas Journal*, must be specified. Also the appropriate quality specifications must be specified in  $b$ . For each activity vector, the operating costs of processing a barrel of input must be stated in the  $c$  vector. These include the cost of crude, natural gas liquids, catalysts, fuel, water, royalty, etc. With the exception of crude costs, these cost data were taken directly from the Bonner and Moore model. By-products were treated as negative inputs and appear as cost credits to reduce the costs of making the basic product mix.

To illustrate the single product case, the marginal cost curve is derived based on the actual product mix and process configurations for 1968. Figure 1 confirms the classical textbook shape of marginal cost curves as it rises over a broad range of output.<sup>10</sup> With actual output in 1968 of

<sup>10</sup> First, the cost minimizing solution for an output level of 8,000,000 barrels per day of the actual product mix was derived. Subsequently, the output constraint was increased by increments from 8,000,000 to 11,000,00 barrels per day of the fixed product mix with each parametric step, obtaining the cost minimizing solution at each step. The marginal cost at each output level can be found by the shadow price (dual solution) corresponding to the output constraint or by merely (over)

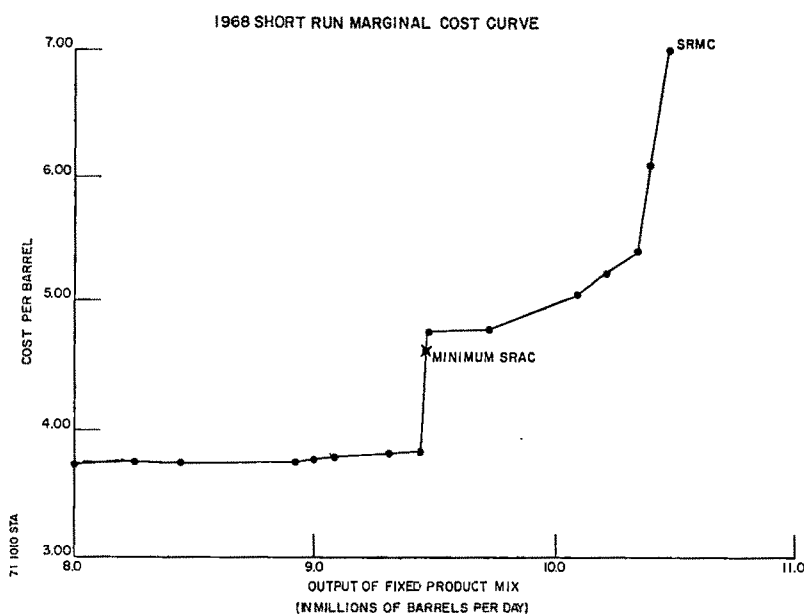


FIGURE 1

9,540,000 barrels per day (9.5 *MMB/D*) of the 1968 product mix, it is clear that the rising part of the marginal cost curve applies to the relevant output range. Also, the minimum of the short-run average cost curve<sup>11</sup> occurs at an output level well below that of maximum output. Unlike the statistical cost function approach, with process analysis the complete output range of the cost function can be investigated rather than only observed cost-output relationships.

Consider the factors underlying such a rising marginal cost curve. First of all, a marginal cost curve with a nonnegative slope follows directly from the concavity assumptions implicit in the *LP* description

dividing the change in costs by the change in output from the previous parametric step. Since over the interval of the parametric steps several basis changes can occur, the marginal cost curve between observations was approximated by linear interpolation. In reality, the marginal cost function would increase in step functions, with each basis change.

<sup>11</sup> Fixed costs were assumed to be \$10,870,000 per day in 1968. See Griffin, Appendix B.

of the production function. However, such concavity assumptions could give rise to a marginal cost curve with an inverted L shape. This shape would imply that marginal costs were flat up to a physical limit on output and that average costs decline over the total output range.

Rather the phenomena producing the upward sloping marginal cost curve can be traced back to the limited capacities of the twelve major process units. Table 1 shows how the increased marginal cost is associated with the various processes reaching capacity.

At the maximum physical output of 10.5 *MMB/D*, most of the processes have reached their individual engineering capacity. It is interesting to note that process analysis will show those capital processes which are redundant for all levels of output such as isomerization. Some units need not approach capacity as shown by the activity levels of catalytic reforming and alkylation. The activities of these units are reduced at very high output levels be-

TABLE 1—RELATIONSHIP BETWEEN MARGINAL COSTS AND  
PROCESS CAPACITY UTILIZATIONS

Output Level ( <i>MMB/D</i> )	8.00	9.00	9.72	10.50
Marginal Cost (\$/B)	3.72	3.78	4.74	8.72
Process Utilization Rates (%):				
Atmospheric distillation	74.8	85.6	94.4	100.0
Vacuum distillation	81.9	94.6	100.0	100.0
Hydrocracking	100.0	100.0	100.0	100.0
Catalytic cracking	84.5	100.0	100.0	100.0
Thermal cracking	0.0	0.0	9.2	28.2
Thermal reforming	0.0	100.0	100.0	100.0
Catalytic reforming	66.1	72.4	89.4	75.1
Alkylation	87.8	91.6	97.3	87.1
Catalytic polymerization	27.8	60.3	46.2	76.3
Isomerization	0.0	0.0	0.0	0.0
Coker	100.0	100.0	100.0	100.0
Visbreaker	100.0	100.0	100.0	100.0

cause they consume inputs which are more valuable in other uses.

When a particular process unit reaches capacity, larger outputs can still be produced through substitution between processes. But such substitution involves a cost. For example, in making gasoline, the catalytic cracking unit operates most efficiently on light gas oils but can process a much heavier feedstock called light vacuum tower bottoms. The older thermal crackers can operate only on the light gas oils. Once capacity is reached in the catalytic cracker, the thermal crackers are then used to process the light gas oil feed. The catalytic cracker must then be used to process almost entirely the light vacuum tower bottoms. Since light gas oils are processed by the more expensive thermal cracking units, marginal costs rise.

In summary, for the single product case, process analysis enjoys three relative advantages over the statistical cost function approach. First, with a process analysis approach, changes in costs can be linked directly with the limited capacity of the process units and the substitution between the various processes. Alternatively, statistical cost function studies using aggregate capital series to allow for capital changes assume, in effect, that all capital goods

have an infinite elasticity of substitution with each other. Second, with process analysis the complete range of the cost curve can be investigated rather than a limited range of actual observations. Third, the effect of technological change as reflected in new processes and product mixes can be explicitly built into the cost function.<sup>12</sup>

### III. The Joint Product Case

Process analysis is particularly adaptable to the problem of deriving the joint product cost function because it explicitly focuses on the relative product prices which affect the product spectrum. After examining the mathematical formulation of the problem, the industry's joint product cost function will be derived on the basis of 1968 product prices. Subsequently, it is shown that changes in relative product prices for gasoline, kerosene, and distillate yield substantially different joint product cost functions. The effect of changing product prices further points out the deficiencies of the aggregate statistical cost function approach. Since in addition to aggregating over processes as in the

<sup>12</sup> For examples of the effect of variations in the product mix on the cost function, see Griffin.

TABLE 2—1968 VARIABLE COST FUNCTION<sup>a</sup>

Quantity Responses	Parametric Steps					
	1	2	3	4	5	6
Revenue ( <i>MM\$/D</i> )	34.60	40.76	43.63	47.23	49.94	53.02
Operating Costs ( <i>MM\$/D</i> )	25.00	30.00	32.32	35.24	38.03	41.43
Profit ( <i>MM\$/D</i> )	9.40	10.76	11.31	11.99	11.91	11.59
Product Outputs: ( <i>MMB/D</i> )						
Gasoline	4.14	4.73	5.02	5.37	5.51	5.83
Kerosene (+ Kerosene Jet)	1.24	1.46	1.57	1.70	1.64	1.57
Distillate	.81	1.07	1.18	1.33	1.72	2.10
Residual Fuel	.31	.53	.61	.71	.72	.73
Asphalt, Road Oil & Lube Stock	.14	.24	.31	.40	.73	1.06
Military Jet	—	—	—	—	—	—
Still Gas	.30	.33	.35	.37	.39	.34
LPG for Fuel & Chemicals	.25	.26	.25	.25	.26	.22
Petrochemical Feedstocks	.23	.28	.29	.32	.34	.31
Coke	.25	.26	.27	.28	.24	.20

<sup>a</sup> See the Appendix for product prices.

single product case, the statistical cost function approach also aggregates over outputs in the joint product case.

The process analysis statement for the derivation of the joint product cost function is to maximize revenue for various rates of constrained operating costs ( $c_0^i, i=1 \dots z$ ) as follows:

(6) maximize  $p'x$

subject to  $c'x^i = c_0^i$  for  $i = 1 \dots z$

$$Ax^i \leq b$$

where  $c_0^i = c_0 + i\delta$

and  $b_j = 0$

The vector  $p$  gives the prices for the respective product outputs, and the outputs are then determined by the *LP* solution.

The joint product cost curve was derived for 1968 using the wholesale market prices for 1968 to represent the marginal revenues for each product. Also the appropriate factor prices, product specifications, and process configurations were used. First, revenue was maximized for an operating cost level of \$25,000,000 per day. At each successive parametric step, operating costs were increased by increments and the revenue maximizing set of product out-

puts was determined for that level of operating costs. Table 2 demonstrates the  $n$  dimensional joint product cost curve, showing the product outputs for each level of operating costs.

Not all product outputs increase as the operating level is increased, so it is apparent that the product mix is not in fact fixed as in the single product case. Figure 2 shows the marginal cost curve and the minimum of the short-run average cost curve calculated on an aggregate value basis. The vertical axis is measured as a pure number since it is interpreted as the dollar cost to produce a dollar's worth of petroleum output. Short-run average costs are calculated as follows:

$$(7) \quad SRAC = \frac{c'x^i + FC}{p'x^i} \quad i = 1 \dots z$$

On a value basis, marginal costs are determined by the reciprocal of the shadow price on the  $c_0^i$  th constraint.

$$(8) \quad \text{Total Value of Output} = p'x = c_0^i y_0^i + b'y^i$$

$$(9) \quad SRMC = \frac{dc_0^i}{dTVO} = \frac{1}{y_0^i} \quad i = 1 \dots z$$

Just as in the fixed product mix case, the

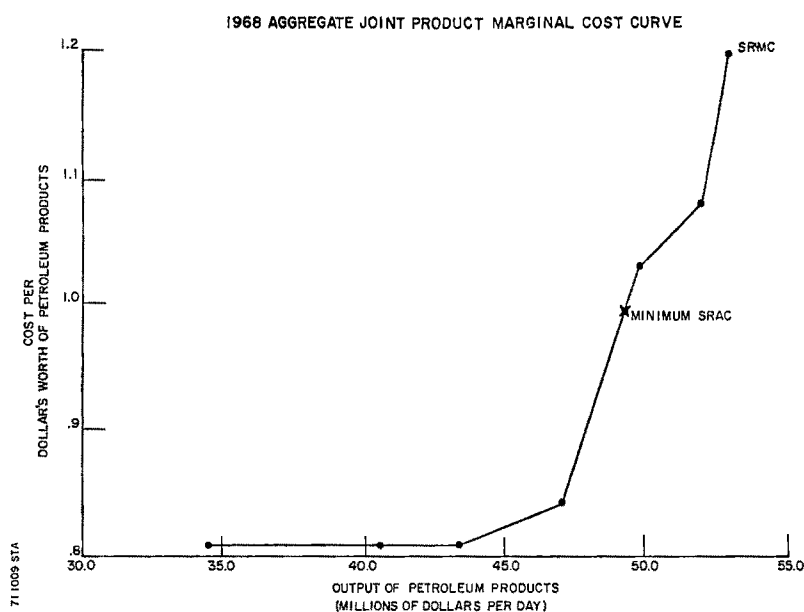


FIGURE 2

aggregate marginal and average short-run cost curves have the classical shape due to the limited capacities of the various processes. As the processes are forced to operate at higher output levels, the refinery produces proportionately greater quantities of lower valued products such as asphalt and distillate fuel. Thus on a value basis, incremental costs increase at a faster rate than the value of output and marginal costs rise.

Variations in relative product prices affect the product spectrum produced. For example, to examine the gasoline output response to a change in gasoline price, a set of cost curves was derived in which the price of gasoline was varied while holding all other prices constant. The various symbols in Figure 3 show the price multiple applied to gasoline prices, e.g., 1.5 means gasoline prices were set at 50 percent above actual prices. Similar calculations were performed in Figures 4 and 5 in which the price of kerosene and distillate were varied, respectively, holding all other prices constant.

The dark solid lines resemble product transformation curves. The one nearest the origin in Figure 3 shows, for an operating cost of 25  $MM\$/D$ , the output combinations available of gasoline and other products. The maximum cost line is a dashed line showing the physical output limit. The dotted lines, resembling output expansion paths, show the revenue maximizing outputs for a given set of relative product prices at various levels of operating costs and are identified by the price variation symbol to which the cost function applies, e.g.,  $2P$ ,  $1.5P$ , etc. The approximate minimum point of the short-run average cost curve is denoted by an  $O$  for each joint product cost curve. It is interesting to note that at  $.5P$  kerosene price the minimum point of the short-run average cost function corresponds to the maximum output level indicating that average costs decline over the complete output range. In all other cases, the short-run average cost function is U-shaped.

Figures 3, 4, and 5 reveal that the product transformation curves are not rec-

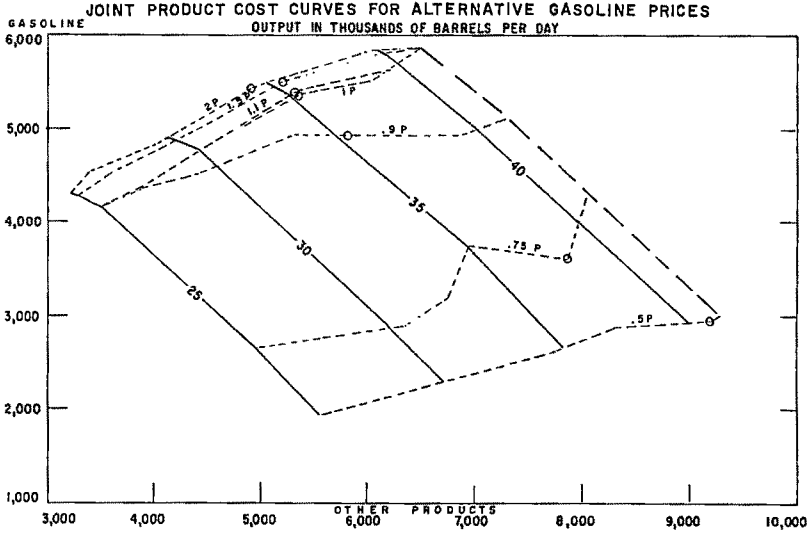


FIGURE 3

tangular, and price changes result in output changes. For gasoline prices, the effect on output is assymetrical. Price increases bring forth little additional output. Yet, price decreases result in substantial substitution of military jet (naphtha type) fuel and distillates for gasoline. At lower gasoline prices, the refiner finds it more valuable to use the naphthas and gas oils

directly for military jet and distillates rather than processing them to make gasoline. For kerosene and distillate price changes, the joint product cost curves shift even more markedly. At  $.5P$  kerosene price, the output of kerosene is zero because it is more profitable to use kerosene to make naphtha jet fuel and distillate fuel oil. Distillate price variations show

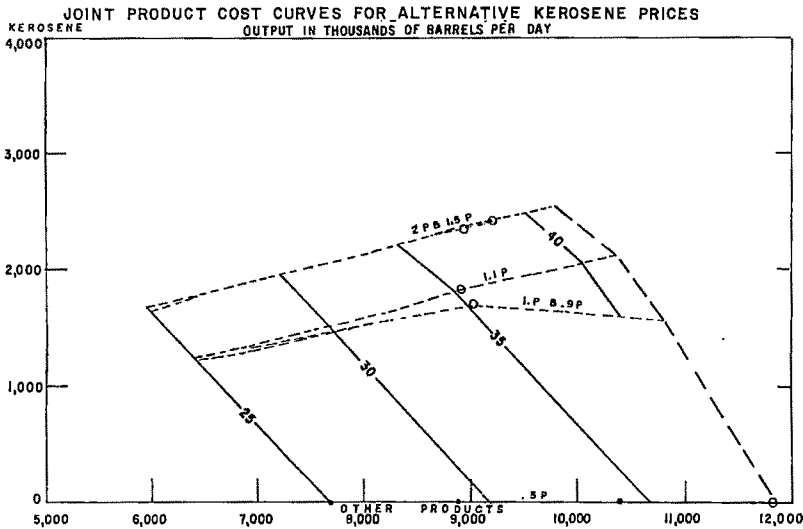


FIGURE 4

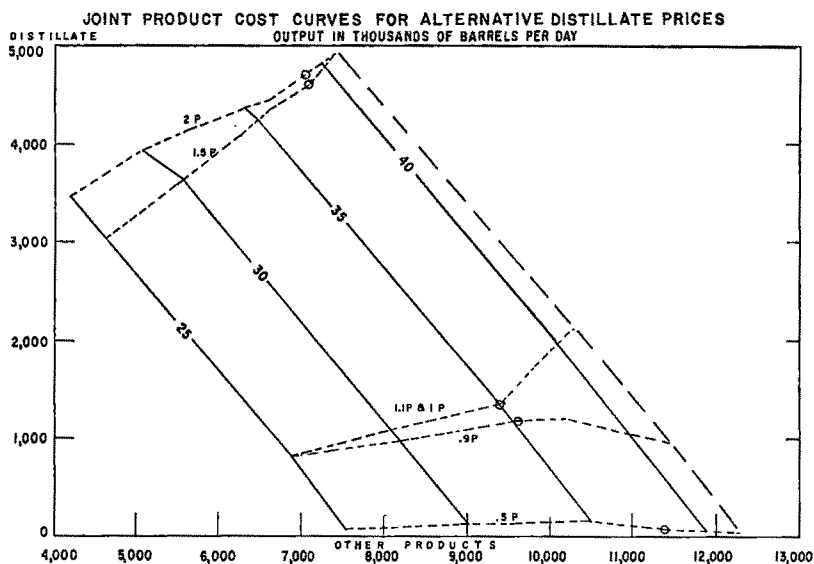


FIGURE 5

that the yield of distillate can vary between 1 percent for  $.5P$  to 50 percent for  $2P$ .

For the joint product case, process analysis enjoys two additional advantages over the statistical cost function approach. First, with the statistical cost function approach, the specific product outputs cannot be derived from the aggregate output index and knowledge of the prices unless the product mix is known or fixed. With process analysis, the specific output level of each product is free to vary and can be explicitly derived as in Table 2. Second, since a change in relative product prices shifts the cost function, the statistical cost function must be based on properly deflated relative factor and product prices. Otherwise, cost-output observations over time will fall along different cost functions. As Johnston points out (see p. 176), the adequacy of the deflation procedure depends on the form of the production function—in fact, there may not be an adequate deflation procedure for product prices as well as factor prices.

#### IV. Conclusions

This paper demonstrates the potential of the process analysis approach for deriving the short-run properties of cost curves. It is useful for both the single and joint product cases. In both cases the results here substantiate the classical assumptions about short-run cost functions, i.e., that marginal costs slope upward and average costs are U-shaped. The rising marginal cost phenomenon can be linked directly to the ability of process analysis to capture the switching between and within process equipment as output expands and impinges on the capacities of the various process units.

#### REFERENCES

- Bonner and Moore Associates, *U.S. Motor Gasoline Economics*, Vol. 1 and 2, Houston 1967, American Petroleum Institute, pub. 4002, 4003. Washington.
- P. J. Dhrymes and B. M. Mitchell, "Estimation of Joint Production Functions," *Econometrica*, Oct. 1969, 37, 732-36.
- G. R. Green, "A Micro-Econometric Analysis of Costs in U.S. Petroleum Refining Estab-

lishments," unpublished doctoral dissertation, Univ. Pennsylvania 1965.

J. M. Griffin, *Capacity Measurement in Petroleum Refining: A Process Analysis Approach to the Joint Product Case*, Lexington, Mass. 1971.

J. Johnston, *Statistical Cost Analysis*, New York 1960.

A. Manne, "A Linear Programming Model of the U.S. Petroleum Refining Industry," *Econometrica*, Jan. 1958, 26, 67-106.

A. Marshall, *Principles of Economics*, New York 1920.

H. D. Vinod, "Econometrics of Joint Production," *Econometrica*, Apr. 1968, 36, 322-36.

A. A. Walters, "Production and Cost Functions: An Econometric Survey," *Econometrica*, Apr. 1963, 31, 1-66.

*Oil and Gas Journal*, Annual Refining Issue, Apr. 1, 1968; Mar. 24, 1969.

*Platt's Oil Price Handbook*, New York 1968.

## APPENDIX

WHOLESALE PRODUCT PRICES USED FOR  
JOINT PRODUCT COST CURVES

	Price <sup>a, c</sup> \$/bbl
Gasoline <sup>b</sup>	5.24
Kerosene & Kerosene Jet	5.05
Distillate	4.13
Residual Fuel	1.81
Asphalt, Road Oil, Lube Stock	1.61
Military Jet	4.64
Still Gas	1.50
LPG for fuel and chemicals	3.15
Petrochemical feedstocks	4.50
Coke	1.00

<sup>a</sup> The prices for regular gasoline, kerosene, distillate and residual fuel are the U.S. wholesale prices published by Independent Petroleum Association of America based on the wholesale low quotations from Platt's Oilgram Price Service.

<sup>b</sup> Gasoline was constrained to be 58 percent regular grade, 40.6 percent premium grade, and 1.4 percent aviation grade. The price was calculated by using the above weights and applying a 2¢/gallon differential for premium grade above regular grade and a 4¢/gallon differential for aviation gasoline above regular grade.

<sup>c</sup> The remaining prices are unpublished, so it was necessary to use a variety of sources and approximations such as Bonner & Moore and corporate estimations. All are wholesale prices.

# Optimal Economic Policy and the Problem of Instrument Instability

By ROBERT S. HOLBROOK\*

It has been said that in a world of perfect knowledge regarding the structure of the economic system and the values of all exogenous variables, policy making becomes a trivial problem.<sup>1</sup> This is true in the sense that it is always possible, with full knowledge and perfect foresight, to choose the appropriate policy for the following period, but it ignores the fact that the correct policy on a period to period basis may turn out to be impractical or even impossible in the longer run. Current policy decisions do not ordinarily have their impact solely in the current period, but rather over a number of periods in the future. Thus, in addition to offsetting the undesired effects of changes in exogenous variables, current policy decisions must offset the current impact of past policy decisions as well. The purpose of this paper is to show that, under quite reasonable assumptions, attempts to offset completely the cumulative impact of past changes in the policy instrument may require ever greater changes in the future value of the instrument, a situation we will characterize as one of "instrument instability." Edward Gramlich has recently discussed the possibility of instrument instability. He suggests that the inclusion of the instrument in the policy maker's welfare function is

the appropriate way to handle the difficulty.

This question of instrument instability has not been of great importance in the past, as little or no attempt has been made to control the behavior of the economy within precise limits. As long as policy makers act slowly, merely nudging the economy in one direction or the other, it is unlikely that instrument instability would even be noticed, much less that it would be a serious problem. Now that they are attempting to "fine tune" the economy, however, the question of instrument instability becomes important; if it is found to exist, serious thought must be devoted to the problem of dealing with it.

This paper describes the nature of the problem, presents two simple cases in detail, outlines the general case, cites some empirical evidence, and suggests some remedies. It shows that the question is really one of the magnitude of the weights on current and lagged values of policy changes as they affect current values of the target or goal variables, and that weighting patterns similar to those frequently found in economic analyses can easily give rise to instrument instability.

## I

We will use a very simple model, consisting of a single relationship comprising a policy instrument  $P$ , an exogenous variable  $X$ , and a goal variable  $Y$ .

$$(1) \quad Y_t = X_t + \sum_{i=0}^n w_i P_{t-i}$$

We define  $X$  to include all current influences on  $Y_t$  other than those

\* Associate professor of economics, University of Michigan. I have benefitted greatly from the comments and suggestions of my colleagues Saul Hymans, Michael Manove, Lester Taylor, and Sidney Winter; of members of the Money Seminar at the University of Michigan; and of an anonymous referee. They should not be held responsible for any errors that remain.

<sup>1</sup> See, for example, Holbrook and Harold Shapiro, William Poole (1970b), and William Brainard. In fairness to these authors, it must be pointed out that they were all working within a short-run framework, while the problem discussed in this paper has to do with policy making over the long run.

contributed by current and lagged values of the instrument, i.e., it includes the net impact of current and lagged values of all policy instruments (other than  $P$ ) and other truly exogenous variables as well as lagged values of endogenous variables. The units in which  $P$  is measured are defined such that  $\sum_{i=0}^n w_i = 1$ , so the long-run multiplier for  $P$  is unity.<sup>2</sup>

We assume that the policy maker knows the precise values of  $X_t$  and the  $w_i$ , that he can control  $P_t$  without error, and that he has some means of selecting a desired value for his goal variable,  $Y_t$ . With complete knowledge and in the absence of any stochastic or unexplained elements,<sup>3</sup> the policy maker can easily select the optimum value for his instrument as in (2).

$$(2) \quad P_t = \frac{1}{w_0} Y_t - \frac{1}{w_0} X_t - \frac{1}{w_0} \sum_{i=1}^n w_i P_{t-i}$$

It is here that earlier analyses have stopped working with this simple model, on the grounds that policy making in such an environment is trivial. They then introduce stochastic elements, incomplete knowledge, or some other complication and proceed to analyze the impact of its inclusion. While it is true that within the framework of decision making for a single period, equation (2) says all we need to know about the optimal value for  $P_t$ , the influence of lagged values of the instrument on the goal variable (instrument instability) can create serious problems for the policy maker in the long run even under the very simple assumptions we are making here.

We might ask whether the fact that an economy is stable in the usual sense does

<sup>2</sup> Equation (1) can be viewed as the reduced form of any of a wide variety of linear structural models of macro-economic behavior. In particular, it could equally well have come from a Keynesian or a Monetarist model of the economy.

<sup>3</sup> The policy maker is assumed to be able to forecast all variables without error; nothing is omitted, to appear later in an error term.

not imply that it will also exhibit instrument stability. Unfortunately, we can readily see that there is no necessary relation between the two types of stability. Let  $Y_T$  be a vector of the current values of the goal variables (i.e., the ones whose values appear in the policy maker's welfare function), while  $Y_L$  is a matrix of all lagged values of these goal variables which have any impact on their current values. Assume equality in the number of goal variables and instruments, and let  $P_T$  and  $P_L$  be, respectively, a vector and a matrix of current and lagged values of the instruments. The behavior of the economy can then be represented by (3),

$$(3) \quad A Y_T + B P_T + C Y_L + D P_L + E Z = 0$$

where  $Z$  is a matrix of all other variables which influence the values of the goal variables, and  $A$  through  $E$  are matrices of coefficients ( $A$  and  $B$  are square and non-singular, by assumption). Solving for  $Y_T$ , we obtain equation (4),

$$(4) \quad \begin{aligned} Y_T = & -A^{-1} B P_T - A^{-1} C Y_L \\ & - A^{-1} D P_L - A^{-1} E Z \end{aligned}$$

and stability in the usual sense is determined by the values of the elements of  $A^{-1}C$ . The question of instrument stability is not answered in equation (4), but rather in equation (5)

$$(5) \quad \begin{aligned} P_T = & -B^{-1} A Y_T - B^{-1} C Y_L \\ & - B^{-1} D P_L - B^{-1} E Z \end{aligned}$$

and depends upon the values of the elements of  $B^{-1}D$ , which have no direct connection with those of  $A^{-1}C$ . Thus it would be quite possible for an economy to display stability in one sense and instability in the other, and both are matters of importance to a policy maker.

Returning to our initial model, we can simplify further by putting equation (2) into first difference form and assuming, without loss of generality, that the policy

maker desires to fix the value of his goal variable at a constant value; this yields equation (6), where  $\Delta Y_t$  is set equal to zero.

$$(6) \quad \Delta P_t = -\frac{1}{w_0} \Delta X_t - \frac{1}{w_0} \sum_{i=1}^n w_i \Delta P_{t-i}$$

Let us now assume that, starting from equilibrium, where  $\Delta P = \Delta X = 0$ ,  $\Delta X_t$  takes on some nonzero value for a single period, and then returns to zero. The question that must be answered is whether the required path of future changes in the instrument converges to zero, or whether it diverges from zero, requiring ever greater policy maneuvers in order to maintain a fixed value for the goal variable. The answer clearly depends on the values of the  $w_i$ , but before moving to the general solution, we will first examine a special case.

A very simple weighting scheme encountered frequently in economics is the geometrically declining series  $b$ ,  $b(1-b)$ ,  $b(1-b)^2$ , etc., where  $0 < b < 1$ . We can examine the stability implications of this set of weights by making the substitution as shown in equation (7),

$$(7) \quad \Delta P_t = -\frac{1}{b} \Delta X_t - \sum_{i=1}^{\infty} (1-b)^i \Delta P_{t-i}$$

and then using the standard transformation to obtain (8).

$$(8) \quad \Delta P_t = -\frac{1}{b} \Delta X_t + \frac{1-b}{b} \Delta X_{t-1}$$

Equation (8) clearly shows that in order to offset the impact on the goal variable of a once for all change in  $X$ , the policy maker need only make two adjustments in the policy instrument, one during the period when  $X_t$  changes, and one the following period. Thus geometrically declining weights would be very convenient. Nevertheless empirical research with less restrictive weighting functions reveals that a more common result is for the weights to

start small, rise to a peak, and then decline, at times becoming negative. Such a set of "humped" weights does not yield so simple a result as the geometrically declining ones, and their implications are not as easily derived.

Simplifying the problem further, we note that the question of instrument instability can be put in the following way: starting from any arbitrary set of past values of the instrument variable, does the implied future path of the instrument converge to a stable value? Thus we can ignore  $\Delta X_t$ , and rewrite the equation as in (9) or (10).

$$(9) \quad \Delta P_t = -\frac{1}{w_0} \sum_{i=1}^n w_i \Delta P_{t-i}$$

$$(10) \quad w_0 \Delta P_t + w_1 \Delta P_{t-1} + \dots + w_n \Delta P_{t-n} = 0$$

In this final form it is revealed as merely a difference equation, and it is a well established result that an equation of this form is stable if, and only if, the roots of the characteristic equation (11) all lie within the unit circle (see Alpha Chiang).

$$(11) \quad w_0 x^n + w_1 x^{n-1} + \dots + w_{n-1} x + w_n = 0$$

Unfortunately, there is a simple general solution only for the case where  $-w_1/w_0 > -w_2/w_0 > -w_3/w_0 > \dots > -w_n/w_0 > 0$ . If this condition holds, the polynomial (11) has a single positive root between 1 and  $\sum_{i=1}^n w_i/w_0$  and this root is the one with the greatest modulus (see K. Sydsæter and Ryuzo Sato). Since we know  $\sum_{i=0}^n w_i = 1$ , it is easy to see that  $\sum_{i=1}^n w_i/w_0 = 1 - 1/w_0$ . Thus, in the case where the only positive weight is attached to the current value of the instrument, and where the past weights decline monotonically to zero, the system is stable. This situation is not commonly encountered in this context, and we must ordinarily expect to have to examine each case individually. For this purpose we can use the Schur theorem (see Chiang) which states that the system will exhibit instrument stability if,

and only if, each of a set of  $n$  determinants composed of ordered arrays of the  $w_i$ 's is positive.

This theorem provides an answer to the stability question regardless of the complexity of the weighting scheme, but it is difficult to draw any simple or general conclusions from it regarding the patterns of weights which would or would not lead to instrument stability. We will instead examine the special cases of one and two lags, and then look at the actual weights that appear in some models whose form resembles that of equation (1). Suppose that there is only a one-period lag, so all  $w_i$  for  $i > 1$  are zero and (11) reduces to  $w_0x + w_1 = 0$ , or  $x = -w_1/w_0$ . Stability requires that  $|x| < 1$ , or  $|w_1| < |w_0|$ . If we combine this requirement with our constraint that  $w_0 + w_1 = 1$ , we can easily see that  $w_0$  must be greater than .5 for stability. If  $w_0$  is less than .5, the policy maker would forever be trying to offset the effect of one period's policy change with a policy change of greater magnitude the following period. Thus, even when the lag in the impact of the instrument is quite short, if the bulk of the impact is not felt until the following period, an attempt to use that tool as a means of maintaining absolute stability in the goal variable will lead eventually to unacceptable changes in the value of the instrument.

The case most frequently encountered in empirical research is the one in which the weights have the same sign, and we will examine that case more closely. Suppose we have the situation depicted in (12)

$$(12) \quad \Delta Y_t = \Delta X_t + w_0 P_t + w_1 \Delta P_{t-1}$$

where  $w_0 < .5$ , implying instrument instability as we have seen. We can then ask: if the policy maker knows that perfect achievement of his goal through the use of his instrument will result in instrument instability, must he avoid using the instrument entirely? The answer is that he can

use it, but he may have to accept some fluctuations in his goal variable in the short run in exchange for reduced fluctuations in his instrument variable in the long run. For example, he could set  $\Delta P_t = -\Delta X_t$ , and the fact that the weights sum to unity will insure that at the end of the second period, the goal variable will be back on its desired path. By taking a two-period horizon the policy maker thus avoids the stability problem entirely.<sup>4</sup>

The preceding is a particular example of a tactic which the policy maker can use if the problem of instrument instability is found to be present. Instead of changing the instrument by the full amount necessary to achieve the desired value for the target, the policy maker can change the instrument by some constant fraction ( $u$ ) of that amount. The use of this device will limit the required changes in the instrument, but only at the cost of some variation in the value of the target variable. The behavior of this system over time clearly depends on the size of  $u$  relative to that of  $w_0$ .<sup>5</sup> The device in the preceding paragraph was an example of this procedure with  $u = w_0$ . Any choice of  $u < 2w_0$  will produce stability in both the goal and the instrument, while a value of  $u$  between  $2w_0$  and unity will produce explosive cycles in both

<sup>4</sup> This result is not peculiar to the two-period case, but is true in the general case as well. On the basis of the general form of (12), namely

$$\Delta Y_t = \Delta X_t + w_0 \Delta P_t + w_1 \Delta P_{t-1} + \dots + w_n \Delta P_{t-n}$$

we can follow the policy of setting  $\Delta P_t = -\Delta X_t$ , and be sure that the impact on  $Y$  of a given change in  $X$  in period  $t$  will have been eliminated by period  $t+n+1$ . Thus the policy maker will always be able to maintain "stability" in his goal variable if he is willing to define the length of the time period to be equal to or greater than the length of the longest lag. While in a theoretical model the length of a period is arbitrary, its choice in the context of this paper is that time interval within which we are willing to ignore variations in the goal variables (i.e., to be concerned only with their average values).

<sup>5</sup> The behavior of the system under different values of  $u$  and  $w_0$  can be shown to be related to the behavior of  $(1 - u/w_0)^n$  as  $n$  becomes large.

variables. Thus, if the long-run behavior of the instrument is of no concern to the policy maker, he should make a full adjustment ( $u=1$ ).<sup>6</sup> If he is concerned about both the goal variable and the instrument, he should choose a value for  $u$  somewhere between 0 and  $2w_0$ .<sup>7</sup>

An alternative approach to the problem of instability when  $w_0 < .5$  is to let  $u=1$  whenever possible, but to put an absolute limit on the permissible change in the instrument during a single period. Although such a rule might seem to have some intuitive appeal, it fails to produce the desired result. Under this regime, as long as the required change in the instrument is less than the maximum allowable change, the goal variable is fixed. As soon as the constraint becomes binding (as must inevitably occur), however, both the goal variable and the instrument exhibit cycles of constant amplitude, and the cycles of the goal variable do not even necessarily center on its equilibrium value.

There are many ways in which a policy choice could be made in the unstable two-period lag case, but all will require some sort of compromise between goal and instrument stability. Within the hypothesized structure, if  $w_0 < .5$ , there is no way in which absolute stability in the goal can be achieved without ever increasing changes in the instrument.

Moving on to the case of two lags, such that all  $w_i=0$  for  $i>2$ , Paul Samuelson gives the necessary and sufficient conditions for instrument stability, as shown in (13),

$$(13a) \quad 1 - \frac{w_2}{w_0} > 0$$

<sup>6</sup> A choice of  $u=1$  implies not only that the instrument fails to appear in the policy maker's utility function, but also that there are no binding constraints on the values it may assume.

<sup>7</sup> A choice of  $u$  between  $2w_0$  and 1 produces the worst of both worlds, and would never be selected by a rational policy maker.

$$(13b) \quad 1 + \frac{w_1}{w_0} + \frac{w_2}{w_0} > 0$$

$$(13c) \quad 1 - \frac{w_1}{w_0} + \frac{w_2}{w_0} > 0$$

and from these and our specification that  $w_0 + w_1 + w_2 = 1$  we can derive<sup>8</sup> the conditions on the individual weights given in (14).

$$(14a) \quad w_0 > .25$$

$$(14b) \quad w_1 < .5$$

$$(14c) \quad w_2 < w_0$$

These conditions are clearly satisfied by any set of declining weights, and they will also be satisfied by weights which first rise and then fall provided that no more than half the weight is on last period's policy, and the impact of policy two periods ago does not exceed the impact of current policy.

Presumably here, too, there is a set of rules for partial adjustment which could turn an unstable situation into a stable one at the cost of some variation in the goal variable, as was the case for the two-period example. Although I have not yet examined this problem closely, it seems reasonable to assume that the same general result would hold, i.e., for zero adjustment in the instrument we lose all control over the goal; for a small adjustment coefficient both goal and instrument exhibit long-run stability; for a full adjustment the instrument is unstable.

## II

Rather than explore further the implications of certain special cases, we turn

<sup>8</sup> The derivation is as follows: If  $w_0 < 0$ , then from (13b) we obtain  $w_0 + w_1 + w_2 < 0$ , which contradicts our specification that the weights sum to one; therefore  $w_0 > 0$ . This result, together with (13a), yields the condition that  $w_0 > w_2$ . Using the fact that  $w_0 > 0$ , and adding  $2w_1$  to both sides of (13c), we obtain  $w_1 < .5$ . Having derived both (14b) and (14c) we add them together, add  $w_0$  to both sides, and obtain  $w_0 > .25$ .

now to the available empirical evidence on the existence of instrument instability. Our equation (1) can be viewed as a linearized reduced form of any structural model of the economy, and our weights  $w_i$  are implicit in the structural coefficients of that model.<sup>9</sup> The reduced forms of such models cannot be easily derived, and their stability characteristics would be most conveniently explored through simulation techniques. There is one exception to this, however, in that Leonall Anderson and Jerry Jordan (A-J) have developed a reduced form model of the U.S. economy which differs from our theoretical one only in that it includes more than a single instrument (generally one monetary and one fiscal instrument, but occasionally more than one of each) and has no representation of our  $X$  variable. Without attempting to imply either approval or disapproval of the A-J approach, we can examine the weighting patterns which appear in some of the examples of that model for their stability characteristics.

There are more than two dozen published versions of the A-J equation, utilizing a wide variety of independent variables, and fitted over many different time periods (Frank de Leeuw and John Kalchbrenner, Edward Corrigan). While it would be possible to examine them all for their stability characteristics, the outcome tends to be repetitious, and a small sample will be sufficient to convey the

results. In a recent version Anderson and Keith Carlson use the quarterly dollar changes in the money stock ( $\Delta M$ ) and full employment federal expenditures ( $\Delta E$ ) as determinants of quarterly changes in the current dollar value of *GNP* ( $\Delta Y$ ) with the results shown in (15).

$$(15) \quad \Delta Y_t = 2.67 + \sum_{i=0}^4 m_i \Delta M_{t-i} + \sum_{i=0}^4 e_i \Delta E_{t-i}$$

where	$m_0 = 1.22$	$e_0 = .56$
	$m_1 = 1.80$	$e_1 = .45$
	$m_2 = 1.62$	$e_2 = .01$
	$m_3 = .87$	$e_3 = -.43$
	$m_4 = .06$	$e_4 = -.54$

$$\sum_{i=0}^4 m_i = 5.57 \quad \sum_{i=0}^4 e_i = .05$$

Applying the Schur theorem to the implied weights we find that while both  $\Delta M$  and  $\Delta E$  are stable instruments,  $\Delta E$  is very nearly unstable (note that  $e_4$  is almost as large as  $e_0$  in absolute value). A small change in one of the weights (e.g., if  $e_4$  were  $-.57$  rather than  $-.54$ ) would be sufficient to tip the balance toward instability. The money supply is a much more stable instrument here. This result is similar to those implicit in the earlier versions of the A-J equation. In those equations,  $\Delta E$  is actually an unstable instrument in most cases, while  $\Delta M$  is uniformly stable. Full employment federal receipts also appear as a stable instrument in several of the equations, and the monetary base appears as an unstable instrument in one of them.

It would be inappropriate to conclude from these results that the money supply is a stable policy instrument while federal expenditures is an unstable instrument. The evidence can only be taken as sugges-

<sup>9</sup> While the linear form may be a satisfactory tool for examination of the behavior of the model within a narrowly defined region, it is likely to be inappropriate for full analysis of instrument instability. Even if the linearized weights imply instability, these weights may in fact be functions of the amplitude of the changes in the instrument, or of other variables within the system which would be affected by the policy choices. The problem may be amplified or diminished by these interactions, but there is no way to include them in the linearized reduced form. Simulation is thus the only analytical tool likely to lead to acceptable answers regarding the prevalence of instrument instability in an economy (unless one is willing to assume that the structure of the economy is in fact linear).

tive, since the A-J approach is but one of many ways of trying to capture the impact of policy instruments on the goal variables. Other models will have to be tested for instrument instability before we can conclude with any degree of certainty that a particular policy instrument is either stable or unstable within the present structure of the *U.S.* economy. Our analysis of the A-J model does indicate, however, that there is a very real possibility that instrument instability could render the precise achievement of goal stability in the United States impossible over the long run.

### III

We have seen that even an omniscient policy maker can be faced with a serious dilemma due to the cumulative effects of past policy choices. The long-run instability of his instrument variable within a regime of full goal achievement may require him to compromise, permitting some goal variability in return for instrument stability.

Far from omniscient, the real world policy maker is in firm command of little if any perfect information, so one might legitimately question the relevance of our conclusions. While it is true that the policy maker must work with imperfect information and rough forecasting models, much current research is designed to improve the quality of our forecasting techniques, and to increase both the quantity and quality of the input data for those forecasts. This research is in part a response to the wide acceptance of the notion that the economy can and should be "finely tuned" to follow precisely the desired path by appropriate adjustments of the policy instruments. Thus, while the problem discussed in this paper may have been academic in the past, it is becoming a question of increasing importance and relevance as our control and forecasting techniques im-

prove.<sup>10</sup> The possibility of instrument instability has not been generally recognized before because policy maneuvers were so grossly in error that the goal variables were highly unstable. The increasing stability of the goal variables (an officially endorsed aim of national policy since World War II, and one which economists have been working to accomplish) may turn out to be unattainable without ever wider and eventually unacceptable fluctuations in some policy instruments.

There are several possible responses to instrument instability, if it should be found to exist. The simplest but perhaps least acceptable approach would be merely to lengthen the period over which we are attempting to stabilize the goal variables. For example, in the one-period lag case discussed earlier, all stability problems disappear if we define a new period equal to two of the old ones, and then tell the policy maker to maintain goal stability on the average over the newly defined period. It would not ordinarily be necessary to define the new period to be so long as to reduce all lagged weights to zero (such a redefinition will always create stability), but merely to be long enough to incorporate the larger lagged weights into the newly defined "current" period. This is an artificial solution, of course, as, in some absolute sense, the goal variable is only apparently stable.

Another technique which bears considerable resemblance to current economic policy decisions in the United States is to choose today's policy so as to achieve the desired goal at some time in the future, on the grounds that today's policy has little impact today, but a growing impact with the passage of time. With the correct choice of the future period for which cur-

<sup>10</sup> Even in the absence of improved control techniques, the problem of instrument instability could arise if the policy maker relied too heavily on a forecasting model which exhibited those characteristics.

rent policy is designed, the instrument stability problem can be overcome, but only at the expense of incomplete goal achievement. Unless the impact of current policy on the current value of the goal variable is actually zero, the continual adjustment of current policy to meet future needs will, in general, prevent the achievement of the goal in the current period. This suggestion, as well as the previous one, touches on the problem of the optimum forecast period, a topic which has not received the attention it deserves.

The policy maker can also avoid the stability problem by making only a partial rather than a full adjustment toward the value of the instrument necessary to maintain complete goal stability. This was demonstrated earlier in the one period lag case, and an appropriate choice of the partial adjustment coefficient would preserve instrument stability in the general case as well.<sup>11</sup>

Finally, it might be that some instruments exhibit instability, while others do not. In this case it would be possible to avoid the problem merely by utilizing the stable instruments. If we have a multiplicity of goals, however, it is likely that discarding one instrument entirely will take us farther from the social optimum than would the judicious use of all instruments.

Each of these remedies (except the last, and only then if the problem instrument is redundant) purchases instrument stability at the expense of greater variability elsewhere in the system, or merely papers over the problem by redefining the period. A

<sup>11</sup> The replacement of discretionary policy by a rule for those instruments which are unstable would obviously avoid the problem (this is equivalent to the case of  $u=0$  in the example). With a good forecasting model, however, it is possible to do better than this. We can reduce goal variability below that consistent with a rule by choosing some small positive partial adjustment coefficient. Thus, the existence of instrument instability would not eliminate the possibility of discretionary policy, it would only limit its ability to control the economy.

complete cure would require a modification in the structure of the system, increasing the impact of current policy and decreasing the impact of past policy.

Much further analysis will be necessary before we can say for certain whether instrument instability is a problem within the structure of the U.S. economy. The actual reduced forms of such large scale models as the FRB-MIT, DIII, or Wharton would be far more complex than the one represented in (1), and it could be that the complex interactions and nonlinearities present in the real economy are such as to make instrument instability either more or less likely than in the strictly linear case. Experiments with these models are required before we can draw more than tentative conclusions about these issues, but the result of some simulations reported by Poole (1970a) indicate that the money supply in the FRB-MIT model may well be an unstable instrument.

If the resources currently being invested on research in the areas of forecasting and optimal policy are to be of greatest possible benefit, the questions raised in this paper must be fully investigated. Otherwise, we may find ourselves in possession of better forecasting techniques than we know how to use.

#### REFERENCES

- L. Anderson and K. Carlson, "A Monetarist Model for Economic Stabilization," *Fed. Res. Bank St. Louis Rev.*, Apr. 1970, 52, 7-25.
- L. Anderson and J. Jordan, "Monetary and Fiscal Actions: A Test of Their Relative Importance," *Fed. Res. Bank St. Louis Rev.*, Nov. 1968, 50, 11-24.
- W. Brainard, "Uncertainty and the Effectiveness of Policy," *Amer. Econ. Rev. Proc.*, May 1967, 57, 411-25.
- A. Chiang, *Fundamental Methods of Mathematical Economics*, New York 1967.
- E. G. Corrigan, "The Measurement and Im-

- portance of Fiscal Policy Changes," *Fed. Res. Bank New York Mon. Rev.*, June 1970, 52, 133-45.
- R. Davis, "How Much Does Money Matter? A Look at Some Recent Evidence," *Fed. Res. Bank New York Mon. Rev.*, June 1969, 51, 119-31.
- F. de Leeuw and J. Kalchbrenner, "Monetary and Fiscal Actions: A Test of Their Relative Importance in Economic Stabilization—Comment," *Fed. Res. Bank St. Louis Rev.*, Apr. 1969, 51, 6-11.
- E. M. Gramlich, "The Usefulness of Monetary and Fiscal Policy as Discretionary Stabilization Tools," *J. Money, Credit, Banking*, May 1971, 3, 506-32.
- R. Holbrook and H. Shapiro, "The Choice of Optimal Intermediate Economic Targets," *Amer. Econ. Rev. Proc.*, May 1970, 60, 40-46.
- W. Poole, (1970a), "Gradualism: A Mid-Course View," in A. Okun and G. Perry, eds., *Brookings Papers on Economic Activity*: 2, Washington 1970, 271-96.
- , (1970b), "Optimal Choice of Monetary Policy Instruments in a Simple Stochastic Macro Model," *Quart. J. Econ.*, May 1970, 84, 197-216.
- P. Samuelson, *Foundations of Economic Analysis*, New York 1965.
- R. Sato, "A Further Note on a Difference Equation Recurring in Growth Theory," *J. Econ. Theor.*, 1970, 2, 95-102.
- K. Sydsaeter, "Note on a Difference Equation Occurring in Growth Theory," *J. Econ. Theor.*, 1969, 1, 104-06.

# Default Risk, Scale, and The Homemade Leverage Theorem

By VERNON L. SMITH\*

Many proofs of the Modigliani-Miller (M-M) theorem on homemade leverage have been provided under the assumption that (i) the corporation can issue unlimited quantities of bonds and shares without altering the default risk on bonds (see Peter Diamond, Franco Modigliani and Merton Miller, Jan Mossin, Joseph Stiglitz), and (ii) the firm's gross cash flow from investment is independent of the amount invested (Modigliani and Miller, Mossin, Stiglitz). Assumption (i) is extremely limiting, and ignores the fact that if there is any chance that the rate of return per dollar invested by a corporation will be less than the interest rate on its bonds, then there can exist a debt-equity ratio large enough to yield positive probability of default. Also, of course, if there is no default risk on bonds, then bonds of the same maturity issued by different corporations are indistinguishable from each other, from Treasury bonds and insured savings accounts, and one cannot account for the widely differing discounts on corporate bonds nor for the survival of various bond rating financial services. Apart from the differential effects created by corporate taxation, default risk is the *sine qua non* of debt financing. Assumption (ii) limits corporate financial theory to the case in which the scale of the firm is given and independent of capital market considerations. Under the conditions of the

M-M theorem the firm's investment decisions are separable from its financing decisions, and it is not incorrect to take cash flow as given. But in models that admit default risk, it is important to reopen the question of the relationship between the firm's scale and its financing.

An investor's expected utility function will be derived on the assumption that a corporation's bonds may have a positive probability of default, and the corporation's uncertain gross cash flow depends on the capitalization. This last assumption represents a departure from the existing literature in which "Each company . . . has, at the time of trading, effected its investments . . . as well as its financing. . ." (Mossin, p. 750).

Assuming zero default risk on debt, the homemade leverage theorem (Theorem 1) is proved by showing that, in equilibrium, an expected utility maximizer will be indifferent to changes in the corporation's debt-equity ratio (except if he holds no bonds). On the assumption that there is positive default risk on bonds, that the corporate rate of return per dollar invested is independent of the amount invested, and with nonnegativity restrictions on debt and shares, we show (Theorem 2) that the homemade leverage theorem fails by showing that, in equilibrium, an investor will prefer an increase or decrease in the corporate debt-equity ratio according to whether that ratio is greater or less than the investor's ratio of bondholdings to shareholdings. "Prefer" in this context means that the investor's expected utility will increase.

Finally (Theorem 3), it is shown that

\* Professor, department of economics, University of Massachusetts, and visiting professor, Cowles Foundation for Research in Economics, Yale University. I am indebted to the National Science Foundation for research support, and to many readers, seminar participants, and others at Purdue, Yale, and the University of Massachusetts for their comments.

for an unrationed margin purchaser of a corporation's shares, an investor, in equilibrium, will prefer an increase (decrease) in the corporate debt-equity ratio if his personal account borrowing rate is greater (less) than the corporate borrowing rate. It is further shown, that if only the shares are pledged as collateral for the personal loan, a potential creditor to the margin investor will prefer to invest in corporate bonds unless the personal borrowing rate exceeds the bond rate.

The model is one in which all financial decisions are made at the beginning of a single interval of real investment, with the resulting cash flow distributed among investors at the end of the period in accordance with the contractual claim conditions of their securities. We deal initially with a single corporation that can issue one class of debt whose claim on the cash flow is strictly prior to that of a single class of common shares. In addition to these two securities, each investor can hold cash (with zero certain return).<sup>1</sup> Each investor distributes his wealth among the three assets so as to maximize his expected utility of terminal wealth. Generalization to more than one corporation is illustrated by extending Theorem 2 for the case of two corporations.

### I. Returns, Investment, and Wealth

Following M-M, Mossin, and Stiglitz, let  $X$  be the corporation's uncertain cash flow at the end of the period, but it is not assumed that  $X$  is independent of the corporation's capitalization. Instead, let<sup>2</sup>

$$(1) \quad X = F(M, s),$$

<sup>1</sup> The model can be extended to deal with senior and subordinate classes of debt, and with warrant instruments. See my 1970 article for warrant analysis.

<sup>2</sup> See Diamond's article for a formulation relating return to "input," and the state of nature, but which assumes zero probability of default on bonds. Diamond, however, does not regard "input" to be part of the firm's equity capital structure.

where  $s$  is a random state variable with probability density  $g(s)$ , and  $M = Y + Z$  is the corporation's total capitalization (=initial market value of the corporation's fully subscribed issue of shares,  $Y$ , plus bonds,  $Z$ ). Equation (1) is essentially a productivity hypothesis: Gross returns are a function of total capital input, given the "state of the world." In fact, if we interpret  $s$  as a random "input," such as "rainfall," or "sunshine," then  $F$  is just the production function when  $X$  and  $M$  are measured in real terms.

The typical investor, with initial wealth  $W_0$  to be distributed among holdings of cash,  $x$ , shares  $y$ , and risky corporate bonds,  $z$ , ends the investment period with different possible values of wealth given by:

$$\begin{aligned} W &= x, & \text{if } X \leq 0 \\ (2) \quad W' &= X(z/Z) + x, & \text{if } 0 < X \leq ZR \\ W'' &= zR + (X - ZR)(y/Y) + x & \text{if } ZR < X \end{aligned}$$

In this development,  $R-1$  is the *contractual* rate of interest on debt  $Z$ , and must be paid *if earned* before the shareholder claim. Bankruptcy is defined as the set of all states  $s$  with the property that  $X \leq 0$ . A nonpositive cash flow means that all investment in bonds and shares is lost, but because of limited liability the investor's holdings of all "other securities," in this case cash,  $x$ , is safe. One might prefer to associate a lump of probability with just the event  $X=0$ , which comes to the same thing. If  $0 < X \leq ZR$ , shareholders receive nothing, and (except when the equality holds), bonds are in default on principal or interest or both. Our investor lays claim to the fraction  $(z/Z)$  of this cash flow. If  $X > ZR$ , the investor's return on bonds is limited by the contractual rate of interest, so he ends the period with corporate bonds worth  $zR$ . Finally, as a

shareholder, he claims the fraction  $(y/Y)$  of all cash flow in excess of  $ZR$ .

## II. Expected Utility Calculations

Let  $U(\cdot)$  be the investor's strictly concave utility of terminal wealth. Given  $M$ , a distribution on states  $s$  induces a distribution on cash flow,  $X$ . If  $F_s = \partial F / \partial s \neq 0$ , from (1) we can write

$$(3) \quad s = \phi(M, X),$$

where  $\phi_M = -(F_M/F_s)$ ,

$$\phi_X = (1/F_s)$$

Hence, expected utility

$$(4) \quad V = \int_{-\infty}^0 U(x)H(M, X)dX \\ + \int_0^{RZ} U(W')H(M, X)dX \\ + \int_{RZ}^{\infty} U(W'')H(M, X)dX,$$

where  $H(M, X)dX = g(s)ds = g[\phi(M, X)]\phi_X dX$ , with  $ds = \phi_X dX$ . The expression (4) makes clear how expected utility depends on the financial aggregates  $Y, Z, M$ .

If  $\mu = Z/Y$  is the debt-equity ratio of the corporation, then  $M = Z + Y = (1 + \mu)Y = (1 + \mu)(Z/\mu)$ , and

$$(5) \quad W' = [(1 + \mu)zX/\mu M] + x, \\ W'' = zR + [(1 + \mu)yX/M] - \mu Ry + x,$$

Now (4) becomes

$$(6) \quad V = \int_{-\infty}^0 U(x)H(M, X)dX \\ + \int_0^{X^*} U(W')H(M, X)dX \\ + \int_{X^*}^{\infty} U(W'')H(M, X)dX \\ X^* = \mu MR/(1 + \mu)$$

Each investor in the market is assumed to be guided by a criterion function typified by (6). In particular the  $i$ th investor must choose the vector  $(x_i, y_i, z_i)$  given

$(\mu, M, R)$ , the real productive investment function  $F(M, s)$ , and the density  $g(s)$ . The role of management is to choose 1) the real investment activities that determine  $F(M, s)$ , and  $g(s)$ , and 2) the rate  $R$  at which risky debt financing is to be obtained from the market. Imagine the firm issuing a prospectus containing all the information relevant to providing the investor with knowledge of  $(F, g, R)$  which in turn constitute the terms on which residual claim shares and priority claim bonds are issued. But these terms are not sufficient information for the investor to choose a portfolio. He also must know  $\mu = \sum_i z_i / \sum_i y_i$ , and  $M = \sum_i (z_i + y_i)$  which are determined by the aggregate of all investor decisions and represent externality variables inherent in the risk structure of the corporation. It is assumed that the capital market adjustment process is sufficiently sophisticated to provide each investor with knowledge of  $\mu$  and  $M$ , whose values, in equilibrium, must be consistent with the aggregate of individual investor choices. The market aggregates  $\mu, M$  appear in each investor's expected utility because securities in the real world inevitably must represent constrained claims across subsets of the set of all states. This contrasts with the Arrow-Debreu ideal world in which firms finance their investments by issuing security claims conditional upon each state that obtains. Because of transaction and information costs, men do not issue state contingent claims. Instead, they issue priority debt claims, and residual share claims with the result that capital markets are characterized by "externalities,"<sup>3</sup> or interac-

<sup>3</sup> James Quirk long ago treated explicitly the problem of default risk, and was led into the problem of lender interdependence. Quirk's motivation was to study the Kalecki principle of increasing risk, and he dealt with a model of the classical entrepreneur rather than the multiowner limited-liability corporation. But the basic ingredient of any comprehensive theory of corporate finance—default risk—with its corollary investor externalities are contained in his fundamental paper.

tions between investors. In this development the ability of capital markets to issue only bonds and shares is taken as given, although ultimately one would hope to explain this fact in terms of transactions and information cost.

Before using (6) to study the individual investor's ideal portfolio behavior, the effect of assuming stochastic constant returns to scale, defined by the condition that

$$F(M, s) \equiv Mh(s),$$

will be examined. If we let  $\theta$  = net rate of return per dollar invested, then  $X = (1 + \theta)M = Mh(s)$ , and  $\theta = \theta(s) = h(s) - 1$ . Hence, the rate of return per dollar invested is randomly distributed independently of the amount invested,  $M$ . Consequently,  $g(s)ds = f(\theta)d\theta$ , where

$$f(\theta) = g[h^{-1}(1 + \theta)]/h'[h^{-1}(1 + \theta)]$$

and (4) can be put in the form

$$(7) \quad V = \int_{-\infty}^{-1} U(x)f(\theta)d\theta + \int_{-1}^{\theta^*} U(W')f(\theta)d\theta \\ + \int_{\theta^*}^{\infty} U(W'')f(\theta)d\theta$$

where

$$(8) \quad W' = (1 + \mu)(1 + \theta)(z/\mu) + x \\ W'' = zR + [(1 + \mu)(1 + \theta) - \mu R]y + x$$

In (7),  $\theta^* = [R\mu/(1 + \mu)] - 1$  is the default rate of return (Quirk, Smith), and  $V$  is now independent of  $M$ .

As to the importance of default risk to the theory of corporate finance, it should be observed that  $\lim_{\mu \rightarrow \infty} \theta^* = R - 1$ . It follows that if the interest rate on a corporation's bonds is at least as large as the lowest possible rate of real return,  $\theta$ , then  $\mu$  may be large enough to yield a positive probability of default. If  $\theta = -1$  is possible, then default probability

$$\int_{-1}^{\theta^*} f(\theta)d(\theta) > 0$$

for all  $\mu$ . Default risk simply cannot be ignored in any general theory of corporate finance.

### III. The M-M Theorem

Returning to (6), assume the investor selects  $(x^0, y^0, z^0)$  so as to maximize  $V$  subject to  $W_0 = x + y + z$ . The Lagrangian is  $L = V + \lambda(W_0 - x - y - z)$  and Kuhn-Tucker first-order conditions are

$$(9) \quad L_x^0 \leq 0, \quad x^0 L_x^0 = 0$$

$$(10) \quad L_{y^0} = I_1 + \mu I_2 - \lambda \\ = J_1 + (1 + \mu)I_2 - \lambda \leq 0, \\ y^0 L_{y^0} = 0$$

$$(11) \quad L_{z^0} = [(1 + \mu)J_0/\mu] + J_1 - \lambda \leq 0, \\ z^0 L_{z^0} = 0,$$

where it is assumed that the individual purchases are "small" relative to the market, i.e.,  $(\partial\mu/\partial y) = (\partial\mu/\partial z) = (\partial M/\partial y) = (\partial M/\partial z) = 0$ , and we define,

$$I_1 = \int_{x^*}^{\infty} U'(W'')(X/M)HdX, \\ I_2 = \int_{x^*}^{\infty} U'(W'')[(X/M) - R]HdX \\ (12) \quad J_0 = \int_0^{x^*} U'(W')(X/M)HdX, \\ J_1 = \int_{x^*}^{\infty} U'(W'')RHdX$$

where  $I_1 = I_2 + J_1$ . The M-M theorem is concerned with whether a change in  $\mu$  will be offset by private portfolio changes. But what is meant by a "change in  $\mu$ " in a circumstance in which  $\mu$  is not a free parameter controlled by someone? The purpose of exploring changes in  $\mu$  on investor behavior is to provide insight into the functioning of capital markets. Since the

investor is assumed to treat  $\mu$  (and  $M$ ) as given, we imagine an experiment in which such parameters can be varied to determine investor response. (This is no different than varying commodity prices to the consumer in demand theory although such prices are actually determined in the market.) In the present development, however, it will not be assumed that  $(\mu, M)$  are necessarily varied independently. It will be necessary to evaluate  $(\partial V/\partial \mu) = V_\mu$  ( $M$  constant) at  $(x^0, y^0, z^0)$ :

$$(13) \quad V_\mu^0 = - (z^0/\mu^2) J_0 + y^0 I_2$$

Also let  $\partial V/\partial M = V_M$  ( $\mu$  constant) in what follows. The partial  $V_M \equiv 0$  only under stochastic constant returns.

The following M-M "homemade anti-leverage" theorem can be proved:

**THEOREM 1.** *If there is no default risk<sup>4</sup> on corporate bonds, an investor's optimal portfolio will be such that he will never prefer a decrease in the debt-equity ratio, and if  $z^0 > 0$  he will be indifferent to a change in the debt-equity ratio, i.e.,*

$$(dV^0/d\mu) = V_\mu^0 + V_M^0(dM/d\mu) = 0.^5$$

In the absence of default risk no part of the subjective probability mass is below  $X^*$ . Thus  $J_0 = 0$ , and (13) becomes  $V_\mu^0 = y^0 I_2$ . Substituting from (10) and (11),

$$\begin{aligned} (1 + \mu)V_\mu^0 &= y^0 (L_{y^0} - J_1 + \lambda) = y^0 (\lambda - J_1) \\ &= -y^0 L_{z^0} \geq 0, \quad z^0 L_{z^0} = 0 \end{aligned}$$

<sup>4</sup> The theorem fails even under zero default risk on bonds, if the firm's financial structure contains option claims (warrants), (see the Appendix in my 1970 article).

<sup>5</sup> This "envelope" equation assumes in general that  $\mu$  and  $M$  may not vary independently. To see why this is needed, suppose some parameter  $\alpha$  does not enter the given investor's utility function, but does enter the utility of other investors. For example an increase in  $\alpha$  might represent a proportionate increase in the initial wealth of other investors. This alters their purchases of shares and bonds and thus changes  $\mu$  and  $M$  to the given investor. For this investor,  $dV^0/d\alpha$

Hence, if  $z^0 > 0$ ,  $V_\mu^0 = 0$ . The investor may desire an increase in the debt-equity ratio but only if he holds no bonds (and some shares). But why is  $V_\mu^0 = 0$ ? It is because a 1 percent increase in bonds relative to shares can be offset by investors holding exactly 1 percent more bonds relative to shares. This can be seen by differentiating the equilibrium conditions (9)–(11) and solving for  $dy^0/d\mu$  (Smith, p. 458), or by noting that

$$\begin{aligned} \partial(W'' - x)/\partial \mu &= R(dz/d\mu) + y(X/M) - Ry \\ &\quad + (1 + \mu)(X/M)(dy/d\mu) - \mu R(dy/d\mu) \\ &= (1 + \mu)[(X/M) - R](dy/d\mu) \\ &\quad + y[(X/M) - R] = 0 \end{aligned}$$

in (5) implies  $(dy^0/d\mu) = -y^0/(1 + \mu)$ . Therefore, letting  $M = (1 + \mu)Y = (1 + \mu) \sum y_i^0$ , we have

$$\begin{aligned} (dM/d\mu) &= (1 + \mu) \sum_i (dy_i^0/d\mu) + \sum_i y_i^0 \\ &= -[(1 + \mu) \sum_i y_i^0/(1 + \mu)] \\ &\quad + \sum_i y_i^0 = 0, \end{aligned}$$

and finally

$$(dV/d\mu) = V_\mu^0 + V_M^0(dM/d\mu) = 0,$$

which proves the theorem. This holds whether or not  $V_M^0 = 0$ , and therefore does not depend on the assumption of stochastic constant returns to scale. From (6) and (7), however, it is clear that although the investor is indifferent to a change in the debt-equity ratio, he is not indifferent to a change in capitalization except under stochastic constant returns.

In this development, the M-M theorem is interpreted in terms of the invariance of

---


$$= (\partial V^0/\partial \mu)(d\mu/d\alpha) + (\partial V^0/\partial M)(dM/d\alpha), \text{ or } dV^0/d\mu = (dV^0/d\alpha)/(d\mu/d\alpha) = \partial V^0/\partial \mu + (\partial V^0/\partial M)(dM/d\mu) \text{ as in the statement of the theorem.}$$

expected utility, and also of market value, with changes in the debt-equity ratio.

#### IV. A Fundamental Leverage Theorem

We now prove,

**THEOREM 2.** *If a corporation can invest at stochastic constant returns to scale, and the default risk on its bonds is positive, then an investor's optimal portfolio will have the property that he will prefer the corporation to increase, leave unchanged or decrease its debt-equity ratio according as  $\mu \gtrless (z^0/y^0)$ .*

Substituting for  $I_2$  and  $J_0$  from (10) and (11), respectively, and using  $y^0 L_{y^0} = z^0 L_{z^0} = 0$ , equation (13) becomes

$$\begin{aligned} V_{\mu}^0 &= - [(z^0/\mu)(L_{z^0} - J_1 + \lambda)/(1 + \mu)] \\ &\quad + y^0(L_{y^0} - J_1 + \lambda)/(1 + \mu) \\ &= (\lambda - J_1)(y^0 - z^0/\mu)/(1 + \mu) \end{aligned}$$

From the inequality in (11)  $\lambda - J_1 \geq (1 + \mu) \cdot J_0/\mu$ . But with positive default probability on bonds,  $J_0 > 0$ . Since we have stochastic constant returns,  $V_M = 0$ , and

$$\frac{dV^0}{d\mu} = V_{\mu}^0 + V_M^0 \left( \frac{dM}{d\mu} \right) = V_{\mu}^0 \gtrless 0 \text{ as}$$

$$\mu \gtrless \left( \frac{z^0}{y^0} \right)$$

If we define  $1 + z^0/y^0$  as the investor's equilibrium antileverage, with  $(1 + \mu)$  the firm's leverage (taken as given by the investor), then this revealed preference theorem states that if antileverage exceeds (is less than) leverage, the investor would prefer a decrease (increase) in corporate leverage. In the development of the previous section, we no longer have

$$\frac{dy^0}{d\mu} = - \frac{y^0}{(1 + \mu)},$$

and hence

$$\frac{dM}{d\mu} \neq 0^6$$

#### V. A Theorem on Differential Borrowing and Lending Rates

In all of the above it has been assumed that the investor holds some corporate bonds, or in the event of a corner solution in bonds, he does not purchase additional shares by borrowing. Such margin purchases expose the investor to claim liabilities limited only by the personal bankruptcy laws. Therefore, all of the borrower's share assets are legally available to meet the investor's debt obligation. If we let  $x \geq 0$  be the amount borrowed at the contractual rate  $R' - 1$  ( $\neq R - 1$  in general), and assume stochastic constant returns, the investor's terminal wealth is

$$\begin{aligned} W' &= 0, \text{ if } \theta \leq \theta' \\ &= [1/(1 + \mu)] [(x/y)R' + \mu R] - 1 \\ &\geq \theta^*, x \geq 0 \end{aligned}$$

$$W'' = (\theta + 1)y + \mu(1 + \theta - R)y - xR', \text{ if } \theta > \theta'$$

Therefore expected utility is

$$(14) \quad V = \int_{\theta'}^{\infty} U(W'')f(\theta)d\theta$$

The budget constraint is  $W^0 + x = y$ , the Lagrangian  $L = V + \lambda(W^0 + x - y)$ , and first-

<sup>6</sup> The case  $z^0 = \mu y^0$  is an exception, for then we have  $dz^0 = \mu dy^0 + y^0 d\mu$ ; using these equations and writing total differentials for  $W = x$ ,  $W'$  and  $W''$  from (8) gives

$$dW = dx^0$$

$$dW' = dW'' = (1 + \theta)[y^0 d\mu + (1 + \mu)dy^0] + dx^0$$

There will be no change in the investor's wealth contingency set as a result of a change  $d\mu$ , if  $dW = dW' = 0$ . The investor achieves such a portfolio neutralization of the effect of  $d\mu$  if the expression  $y^0 d\mu + (1 + \mu)dy^0$  vanishes, or by the adjustment  $(dy^0/d\mu) = -y^0/(1 + \mu)$ , and hence  $dM/d\mu = 0$ , as in the case of no default risk. Note that we will have  $\mu = z^0/y^0$  for all investors if they have identical utility functions and identical probability assessments (see Stiglitz, p. 790, fn. 12). Also, for this case, it should be observed that constant returns to scale is not required for expected utility to be independent of  $\mu$ . That is, in the proof following Theorem 2, we need not have  $V_M^0 = 0$  to get  $dV^0/d\mu = V_{\mu}^0 = 0$ .

order conditions for an interior maximum (the boundary conditions will not be needed) are:

$$(15) \quad L_{x^0} = -R' \int_{\theta^*}^{\infty} U'(W'') f(\theta) d\theta + \lambda \\ = -R'J + \lambda = 0$$

$$(16) \quad L_{y^0} = \int_{\theta^*}^{\infty} U'(W'') [(1 + \theta) \\ + \mu(1 + \theta - R)] f(\theta) d\theta - \lambda \\ = I_1 + \mu I_2 - \lambda \\ = (1 + \mu)I_2 + RJ - \lambda = 0$$

Also we need

$$(17) \quad V_{\mu}^0 = y^0 I_2$$

We now prove the following theorem showing how divergence between borrowing and lending rates affects investor preference for changes in leverage:

**THEOREM 3.** *Under stochastic constant returns to scale the investor whose optimal portfolio requires him to borrow on personal account to buy shares, and who is not rationed by his creditor, will prefer the corporation to increase, leave unchanged, or decrease its debt-equity ratio according as his personal borrowing rate is greater than, equal to, or less than the lending rate on corporate bonds.*

Substituting from (15) and (16) into (17) yields

$$V_{\mu}^0 = y^0 [\lambda - RJ] / (1 + \mu) \\ = y^0 J(R' - R) / (1 + \mu)$$

since  $J > 0$ ,  $V_{\mu}^0 \geq 0$  as  $R' \geq R$ .

The meaning of the theorem is that a margin investor in shares cannot achieve his most preferred position by personal account borrowing (that is, in equilibrium, he will not be indifferent to a change in the debt-equity ratio) unless he can borrow as much as he wants at the same rate as the

corporation. But he cannot borrow at the same rate by simply pledging his shares against the loan. To see why, we examine the terminal wealth contingencies of a potential creditor of the investor who also has the alternative of buying the corporation's bonds.

1) Consider first the case of a creditor dividing his wealth  $W^*$  between cash,  $W^* - x$ , and a loan,  $x$ , to our margin buyer of shares who pledges the shares as collateral against the loan. The lender's wealth contingencies are:

$$W_1 = \begin{cases} W^* - x, & \text{if } \theta \leq \theta^* \\ W^* - x + [(1 + \theta) \\ + \mu(1 + \theta - R)]y, & \text{if } \theta^* < \theta \leq \theta' \\ W^* - x + R'x, & \text{if } \theta > \theta' \end{cases}$$

These wealth contingencies are represented by the dashed line in Figure 1.

2) Now let the lender divide his wealth between cash,  $W^* - x$ , and a purchase,  $x$ , of the corporation's bonds. The wealth contingencies are now

$$W_2 = \begin{cases} W^* - x, & \text{if } \theta \leq -1 \\ (W^* - x) + (1 + \mu) \\ \cdot (1 + \theta)x/\mu, & \text{if } -1 < \theta \leq \theta^* \\ W^* - x + Rx, & \text{if } \theta > \theta^* \end{cases}$$

and are represented by the solid line in Figure 1.

It is clear that unless  $R' > R$ ,  $W_2$  is at least as good as  $W_1$  in all states,  $\theta$ . Therefore a rational lender will not make a positive loan to a margin buyer, whose sole assets are shares, except at some  $R' > R$ .<sup>7</sup>

<sup>7</sup> This result contrasts with Stiglitz, p. 788, who argues that if the individual margin buyer uses the security as collateral he will be able to borrow at the same rate as the corporation. But due to the priority claim of a corporation's bondholders, a lender will prefer holding bonds to holding the individual's note secured by shares (at the same interest rate) because the bonds yield a better return in all those default states  $-1 < \theta < \theta'$ .

There is a fundamental difference between lending

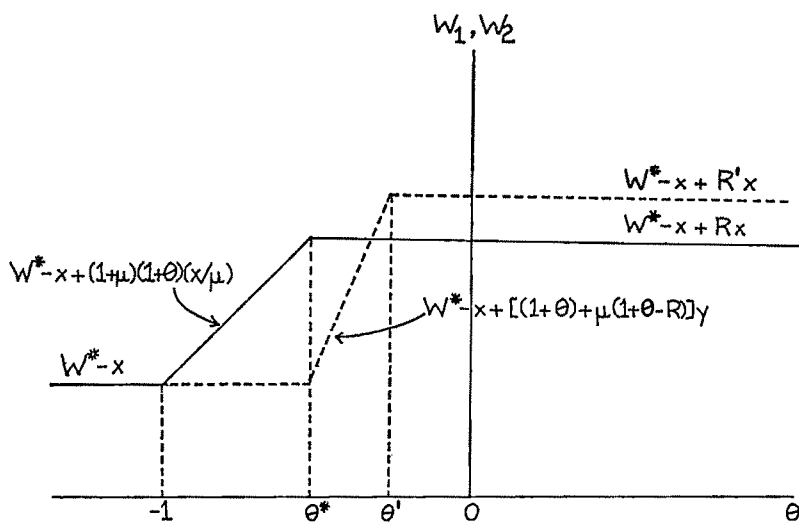


FIGURE 1

## VI. More than One Corporation

In the above we have treated but a single corporation, and the only way that there can be a change in its market value, or capitalization, is for investors to change

to the corporation and lending to an individual, whatever the collateral. Default risk on the corporation's bonds is essentially independent of any one individual's purchase, but default risk on a margin loan depends significantly on the amount of the loan, as is seen by comparing the expressions for  $\theta^*$  and  $\theta'$ .

The point is particularly well illustrated in the case where the individual attempts to reproduce the firm's cash flow contingencies by buying the corporation's bonds and shares in the same ratio that they are issued, then borrowing against this portfolio as collateral. If the individual with wealth  $W_0$  borrows  $x'$ , and invests in  $y$  shares and  $z = \mu y$  bonds ( $W_0 + x' = y + \mu y$ ), his terminal wealth is  $(1 + \mu)(1 + \theta)y - R'x'$ , if  $\theta > \theta'' = [x'R'/(1 + \mu)y] - 1$ , and zero if  $\theta \leq \theta''$ . A lender to this investor who holds  $W^* - x'$  in cash experiences wealth contingencies

$$W_1' = \begin{cases} W^* - x', & \text{if } \theta \leq -1 \\ W^* - x' + (1 + \mu)(1 + \theta)y, & \text{if } -1 < \theta \leq \theta'' \\ W^* - x' + R'x', & \text{if } \theta > \theta'' \end{cases}$$

If he invests  $x'$  in corporate bonds, he ends the period with wealth

$$W_2' = \begin{cases} W^* - x', & \text{if } \theta \leq -1 \\ W^* - x' + (1 + \mu)(1 + \theta)(x'/\mu), & \text{if } -1 < \theta \leq \theta^* \\ W^* - x' + Rx', & \text{if } \theta > \theta^* \end{cases}$$

But  $W_1' = W_2'$  if, and only if,  $R = R'$  and  $x' = \mu y$ , i.e., the amount of the margin loan is exactly equal to the investor's holdings of corporate bonds.

their cash assets. We indicate briefly that the leverage theorem of Section IV holds when there are  $n = 2$  corporations.

Let  $x$  be cash assets,  $y_j$  be shares, and  $z_j$  be bonds,  $j = 1, 2$ . Also let  $\theta_j^* = [R_j \mu_j / (1 + \mu_j)] - 1$  be the default rate of return for corporation  $j$ , and  $\mu_j = Z_j / Y_j$  be the debt-equity ratio for corporation  $j$ ,  $j = 1, 2$ . Then the typical investor's terminal wealth contingencies are:

$$W(\theta_1 \leq -1; \theta_2 \leq -1) = x$$

$$W(-1 < \theta_1 \leq \theta_1^*; \theta_2 \leq -1)$$

$$= x + z_1(\theta_1 + 1)(1 + \mu_1)/\mu_1$$

$$W(\theta_1^* < \theta_1; \theta_2 \leq -1)$$

$$= x + z_1 R_1 + y_1[(\theta_1 + 1) + \mu_1(1 + \theta_1 - R_1)]$$

$$W(\theta_1 \leq -1; -1 < \theta_2 \leq \theta_2^*)$$

$$= x + z_2(\theta_2 + 1)(1 + \mu_2)/\mu_2$$

$$W(-1 < \theta_1 \leq \theta_1^*; -1 < \theta_2 \leq \theta_2^*)$$

$$= x + \sum_{j=1}^2 z_j(\theta_j + 1)(1 + \mu_j)/\mu_j$$

$$W(\theta_1^* < \theta_1; -1 < \theta_2 \leq \theta_2^*)$$

$$= x + z_1 R_1 + y_1[(\theta_1 + 1) + \mu_1(1 + \theta_1 - R_1)] \\ + z_2(\theta_2 + 1)(1 + \mu_2)/\mu_2$$

$$W(\theta_1 \leq -1; \theta_2^* < \theta_2)$$

$$= x + z_2 R_2 + y_2[(\theta_2 + 1) + \mu_2(1 + \theta_2 - R_2)]$$

$$\begin{aligned}
 &W(-1 < \theta_1 \leq \theta_1^*; \quad \theta_2^* < \theta_2) \\
 &= x + z_1(\theta_1 + 1)(1 + \mu_1)/\mu_1 + z_2 R_2 \\
 &\quad + y_2[(\theta_2 + 1) + \mu_2(1 + \theta_2 - R_2)] \\
 &W(\theta_1^* < \theta_1; \quad \theta_2^* < \theta_2) \\
 &= x + \sum_{j=1}^2 \{z_j R_j + y_j[(\theta_j + 1) \\
 &\quad + \mu_j(1 + \theta_j - R_j)]\}
 \end{aligned}$$

The parentheses for each  $W(\cdot)$  indicate how the state space  $(\theta_1, \theta_2)$  is partitioned to get the various combinations of bankruptcy, default, and profitability for  $n=2$  corporations. (The number of such contingencies, and therefore the number of terms in the expected utility calculation, increases as  $3^n$ .) By writing double integrals over these partitions, it is straightforward to write out the expected utility function  $V(x, y_1, y_2, z_1, z_2 | \mu_1, \mu_2, r_1, r_2)$ .

By differentiating the resulting Lagrangian, and substituting as in the proof of Theorem 2, it follows that  $V_{\mu_j}^0 \geq 0$  according as  $\mu_j \geq (z_j^0/y_j^0)$ ;  $j=1, 2$ .

## VII. Financial Management and Default Risk

The corporate finance model of this paper has been employed to answer some questions raised by the M-M theorem. However, the model raises some entirely new questions. The M-M question, "How does the debt-equity ratio affect the cost of capital?", seems to suggest purposive adjustment of  $\mu$  by corporate management. But management cannot directly determine  $\mu$  as this is determined in the aggregate by investors to whom  $\mu$  serves as an externality. To see this, assume that no investor is a margin buyer of shares and we have a tangency solution at  $y_i^0 \geq 0$ ,  $z_i^0 \geq 0$ ,  $x_i^0 \geq 0$  for all investors. Then (9)-(11) become equations defining the individual's equilibrium asset demand functions. For the  $i$ th investor, assuming stochastic constant returns to scale, we have

$$x_i^0(W_{0i}, \mu, R), \quad y_i^0(W_{0i}, \mu, R), \quad z_i^0(W_{0i}, \mu, R)$$

In market equilibrium,

$$\mu \sum_i y_i^0(W_{0i}, \mu, R) - \sum_i z_i^0(W_{0i}, \mu, R) = 0$$

This defines an implicit function,  $F(\mu, R) = 0$  which represents the "finance frontier," a constraint set on management choice of  $\mu, R$ . But in the special case of zero default risk (or under any conditions giving  $\mu = z_i^0/y_i^0$  in Theorem 2),  $\partial y_i^0/\partial \mu = -y_i^0/(1 + \mu) = -\partial z_i^0/\partial \mu$ , as shown above. Therefore,

$$\frac{\partial F}{\partial \mu} = \mu \sum_i \frac{\partial y_i^0}{\partial \mu} + \sum_i y_i^0 - \sum_i \frac{\partial z_i^0}{\partial \mu} = 0,$$

and by the implicit function theorem  $dR/d\mu = -(\partial F/\partial \mu)/(\partial F/\partial R) = 0$  so that  $F(\mu, R)$  is independent of  $\mu$ . Consequently, in this case,  $F(\mu, R) \equiv F(R) = 0$  determines  $R = R^0$ .

But in general, we do not have zero default risk, and we do not have  $\mu = z_i^0/y_i^0$  for every investor. As a result the risky bond rate  $R-1$  is determined in financial markets as an implicit function of  $\mu$ . This gives management freedom to select the terms  $R$  of risky debt finance according to a management criterion function (market value of shares, utility of manager, for example), and via market response, as reflected in  $F(\mu, R) = 0$ ,  $\mu$  is determined. For example, consider the maximization of the market value of shares subject to the above market equilibrium condition. Letting  $\lambda$  be the Lagrange multiplier we maximize the following expression with respect to  $\mu$  and  $R$ :

$$\phi = Y^0 + \lambda(\mu Y^0 - Z^0),$$

subject to

$$Y^0 = \sum_i y_i^0, \quad Z^0 = \sum_i z_i^0$$

giving the following three conditions on  $(\lambda, \mu, R)$ :

$$(18) \quad \frac{\partial Y^0}{\partial R} + \lambda \left( \mu \frac{\partial Y^0}{\partial R} - \frac{\partial Z^0}{\partial R} \right) = 0$$

$$(19) \quad \frac{\partial Y^0}{\partial \mu} + \lambda \left( \mu \frac{\partial Y^0}{\partial \mu} + Y^0 - \frac{\partial Z^0}{\partial \mu} \right) = 0$$

$$(20) \quad \mu Y^0 - Z^0 = 0$$

This development generalizes to  $n$  corporations (see p. 459 of my 1970 article).

By the above interpretation of management's decision role, the assumption of zero default risk on bonds, from which follows the M-M theorem, virtually eliminates the need for financial management (see Mossin, pp. 751, 756). The riskless lending rate is determined by the market, and it is a matter of indifference as to whether firms are financed by debt or equity. So who needs financial management? From whence arises the institution of security underwriting? The answer is intimately bound up with the point made by Stiglitz, p. 792, that "the crucial fallacy lies in the implicit assumption that one firm's bond is identical to another firm's bond." It is also intimately related to the hard facts of externality in the expected utility criterion of each investor. These considerations mean that capital markets have relatively higher information requirements than ordinary Pareto-Walras commodity markets. In general no investor can make a rational allocation of his wealth until he has both price information  $R$ , and equilibrium market information,  $(M, \mu)$ .<sup>8</sup> The function of management in

consultation with financial specialists, such as the underwriter of its debt and equity issues, is to select a financial mix, and an  $R$  that will maximize management's criterion function,<sup>9</sup> and yield a market clearing offer of shares and bonds. This is perhaps the significant function of underwriting. The perfect offering at time zero in our model would specify the number of \$1 share certificates, the volume of bonds, and the contractual bond rate, with both types of securities being exactly subscribed. If any part of the offering is oversubscribed, or if it is undersubscribed and the excess has to be absorbed by the underwriter, then we do not have a cleared market. If either security is oversubscribed on the final offering, then investors can obtain gains from trade and the resulting capital gains represent funds which (in part) could have increased the corporation's capitalization if the offering had been perfect. If either security is undersubscribed and the excess temporarily absorbed under contract by the underwriter, the capital losses of the underwriter (and investors) which results when he eventually unloads the excess offering, indicate that the corporation is overcapitalized relative to a perfect market offering.

The institution of underwriting can be construed as a decentralized response to the externality inherent in the corporate capital market. The underwriter is a specialist in the abnormal information requirements of such a market. Actual offerings are typically made by "preliminary prospectus" which provides the basic financial data together with verbal indica-

<sup>8</sup> With nonconstant returns, conditions (9)-(11) define asset demand functions  $x_i^0(W_{0i}, M, \mu, R)$ ,  $y_i^0(W_{0i}, M, \mu, R)$ ,  $z_i^0(W_{0i}, M, \mu, R)$  for investor  $i$ . In financial market equilibrium there are now two implicit functions in  $(M, \mu, R)$ ,

$$G(M, \mu, R) = M - \sum_i y_i^0 - \sum_i z_i^0 = 0$$

$$H(M, \mu, R) = \mu \sum_i y_i^0 - \sum_i z_i^0 = 0$$

<sup>9</sup> The important point is that management has a degree of freedom that influences the debt-equity menu that confronts investors whatever may be the choice criterion of management. Maximization of market value has been used as an example but, as noted by Stiglitz, p. 790, with default risk there is no reason (i.e., in terms of social welfare) for firms necessarily to maximize market value.

tions of the probable offering price. The offering brokers in the underwriter pool then obtain tentative commitments from their customers. Eventually, the offering is advertised among "next week's probable offerings." A firm price is finally committed, and the offer tendered immediately thereafter. Last minute shading of the price no doubt reflects the result of the preliminary nonbinding commitments as well as general market conditions. This procedure represents an obvious approximation of recontracting, but has the important additional feature of providing investors with a preliminary estimate of the aggregate financing mix. Our model shows they must have this information, as well as price information. The feedback from investors provides information, however crude, on the accuracy of the original estimates. Casual observation of the frequency with which offerings are canceled or postponed and/or prices adjusted would suggest that the procedure has valuable information content. The survival value of underwriting as an institution may suggest that the procedure is relevant

given the risks, and transaction costs, inherent in the process.

#### REFERENCES

- P. Diamond, "The Role of a Stock Market in a General Equilibrium Model with Technological Uncertainty," *Amer. Econ. Rev.*, Sept. 1967, 57, 759-76.
- J. Lintner, "Dividends, Earnings, Leverage, Stock Prices and the Supply of Capital to Corporations," *Rev. Econ. Statist.*, Aug. 1962, 44, 243-69.
- F. Modigliani and M. Miller, "The Cost of Capital, Corporate Finance and the Theory of Investment," *Amer. Econ. Rev.*, June 1958, 48, 261-97.
- J. Mossin, "Security Pricing and Investment Criteria in Competitive Markets," *Amer. Econ. Rev.*, Dec. 1969, 59, 749-56.
- J. P. Quirk, "The Capital Structure of Firms and the Risk of Failure," *Int. Econ. Rev.*, May 1961, 2, 210-28.
- V. L. Smith, "Corporate Financial Theory Under Uncertainty," *Quart. J. Econ.*, Aug. 1970, 84, 451-71.
- J. E. Stiglitz, "A Re-Examination of the Modigliani-Miller Theorem," *Amer. Econ. Rev.*, Dec. 1969, 59, 784-93.

# The Creation of Risk Aversion by Imperfect Capital Markets

By ROBERT TEMPEST MASSON\*

In this paper I advance a rationale for risk-averse behavior by people of little wealth who face imperfect capital markets. In the literature on risk and uncertainty, risk-averse behavior is frequently postulated by assuming a concave utility function (see Irving Fisher and George Hall, and Bernt Stigum), convex indifference curves in a mean-variance model (see Martin Feldstein, Stewart Johnson, and James Tobin), or some "safety-first" criterion (see Jean-Marc Boussard and Michel Petit, and Lester Telser). In a recent article by David Pyle and Stephen Turnovsky, subsets of these criteria for risk-averse behavior are shown under many assumptions to be operationally indistinguishable. On the other hand, the implicit psychological framework underlying these models may differ. In the expected utility or mean-variance framework a psychological aversion to risk is generally implied, whereas under a safety-first criterion the individual often is seen as attempting to avoid illiquidity.

In this paper I offer an analysis which integrates imperfections in the capital markets with expected utility maximization and shows how risk-averse behavior may follow from institutional characteristics of the economy, and not necessarily from a psychological aversion to risk. The advantages of this type of analysis are evident in the policy-oriented hypotheses which it generates. Hypotheses may be

formulated from this model which relate risk-averse behavior to economic institutions and policies.

I present a basic two-period model in which imperfect capital markets are assumed. Intertemporal risk neutrality (to be defined in Section I) is also assumed.<sup>1</sup> Considering an individual's expected utility maximization problem when there are imperfect capital markets (at least one plausible formulation of them), one arrives at the following striking conclusion: *The risk-neutral individual who faces these imperfect capital markets will behave on gambles involving present income as if he were maximizing expected utility on a Friedman-Savage type utility function* (see Milton Friedman and L. J. Savage).

After the presentation of this simple model, some conclusions and familiar examples from the model are sketched out to demonstrate its applicability.

## I. The Basic Model

The first step is to consider the case of a person who is intertemporally risk-neutral for real income and then to indicate

<sup>1</sup>The assumption of intertemporal risk neutrality provides a base point. There is no reason to assume either a risk-loving or a risk-averse psychology. If I had made the assumption of risk aversion, the imperfect capital markets would create even more risk-averse behavior and even more risk-loving behavior for different asset levels than the utility function would imply with perfect capital markets. Joseph Stiglitz implies that the assumption of intertemporal risk neutrality is probably bad because we do not observe straight line Engel curves, p. 667. The rationale for my use of intertemporal risk neutrality is given in the next section. The validity of Stiglitz's inference is, however, affected by his assumption of perfect capital markets.

\* Assistant professor of economics at Northwestern University. A recent article by Nils Hakansson treats a closely related problem using intertemporally additive, risk-averse utility functions.

how imperfect capital markets<sup>2</sup> will make him behave *as if* he were risk-averse at any point in time.

I define intertemporal risk neutrality as follows: Give an individual a choice between 1) having the consumption stream  $(c_0, c_1, \dots, c_T)$  for the time periods 0 (the present) to  $T$  (time of death), and 2) a .5 probability of a consumption stream

$$(c_0(1 + \epsilon), c_1(1 + \epsilon), \dots, c_T(1 + \epsilon))$$

and a .5 probability of a consumption stream

$$(c_0(1 - \epsilon), c_1(1 - \epsilon), \dots, c_T(1 - \epsilon))$$

The risk-neutral individual will be indifferent between the certain stream and the gamble. If a utility function which transforms  $(c_0, \dots, c_T)$  into a von Neumann-Morgenstern type of quasi-cardinal utility function (see William Baumol, pp. 512-16) is specified, and perfect (single interest rate) capital markets are assumed, then this definition may be restated. In this case a person is risk-neutral by the above definition if he is indifferent between an income  $y_i$  or a .5 probability of  $y_i + \epsilon_i$  and of  $y_i - \epsilon_i$  for any year  $i$ . This assumes that the consumption allocations for all periods are decided upon after knowing the results of the gamble.<sup>3</sup>

To demonstrate the process by which

<sup>2</sup> The reader should note that the term imperfect capital markets is used to mean that the interest rate paid or earned is a function of the amount borrowed or lent (see Jack Hirshleifer, p. 329). I am not stating that capital markets are indeed imperfect in the sense of the term used in industrial organization (see George Stigler).

<sup>3</sup> The general proof of this is somewhat longer than is worth presenting here. Just note that the first definition puts us in the class of linear homogeneous utility function. Thus, a shift in a straight line budget constraint caused by an *income* change in any one period moves the *consumption* equilibrium along a ray from the origin (see fn. 5). If a person's consumption decisions are made before knowing the results of the gamble, the stochastic nature of  $\epsilon_i$  is shifted to  $c_i$ , which automatically creates risk aversion since decreasing marginal utility occurs in each time period (see fn. 6).

imperfect capital markets convert the risk-neutral individual to risk-averse behavior in the present, I shall specify a particular structure of capital markets and calculate the individual's first-period behavioristic utility function.

The definition of risk neutrality used here requires a utility function which is linear homogeneous (see Stiglitz) and has a diminishing marginal rate of substitution.<sup>4</sup>

I shall assume that utility is a linear homogeneous function of consumption:

$$(1) \quad u = u(c_1, c_2)$$

where:  $u$  = utility level,

$c_1$  = dollars of consumption in period one, and  
 $c_2$  = dollars of consumption in period two

If there were only a single interest rate and endowment incomes of  $e_1$  and  $e_2$  given exogenously, the problem would be to:

$$(2) \quad \text{maximize } u(c_1, c_2)$$

$$\text{subject to} \quad c_1 = e_1 + b, \\ c_2 = e_2 - \rho b,$$

$$\text{and} \quad b \geq 0$$

where:  $b$  is the money borrowed ( $b > 0$ ) or invested ( $b < 0$ ) measured in time one dollars;  
 $\rho$  is one plus the interest rate

This problem may be rewritten:

$$(3) \quad \text{maximize}_b u(e_1 + b, e_2 - \rho b)$$

and is solvable for a reduced form or real income function:

$$(4) \quad u^* = u^*(e_1, e_2, \rho)$$

This real income function obeys the von Neumann-Morgenstern expected util-

<sup>4</sup> The reader may note that Stiglitz tends to reject the implied homotheticity for empirical reasons. This is due to his assumption of straight line budget sets, p. 667.

ity maximization axioms for either period's income if the other income and  $\rho$  are known.

It may also be solved for the borrowing function:

$$(5) \quad b = b(e_1, e_2, \rho)$$

At this point imperfections in the capital markets may be introduced. This may be done by making  $\rho$  a function of  $b$ :

$$(6) \quad \rho = \rho(b)$$

If money capital becomes more expensive as more is borrowed, then:

$$(7) \quad \rho'(b) > 0$$

for  $b \geq 0$

For simplicity of mathematical form I shall first consider the function  $\rho(b)$  to have the form:

$$(8) \quad \begin{aligned} \rho(b) &= \rho_\beta & \text{if } b \geq 0 \\ \rho(b) &= \rho_\alpha & \text{if } b < 0 \end{aligned}$$

where:  $\rho_\beta$  is one plus the borrowing rate of interest;

$\rho_\alpha$  is one plus the lending rate of interest; and

$$\rho_\beta > \rho_\alpha$$

This simplifies solution because for any  $b$  equal to zero,  $\rho'(b)$  is defined and equal to zero.

The above utility expression may be maximized for any specific  $\rho_\beta$  and  $\rho_\alpha$ . This in fact becomes a programming problem, but it need not be put explicitly into the programming notation. The income consumption curve,  $ICC$ , and the three possible types of solutions may be illustrated graphically.

Figure 1 shows successive equilibria when period two income is held constant at  $\bar{e}_2$  and period one income expands. It exhibits the familiar condition for homogeneous functions that for a single price ratio (interest rate) the income consumption curve ( $ICC$ ) (any isocline) is a

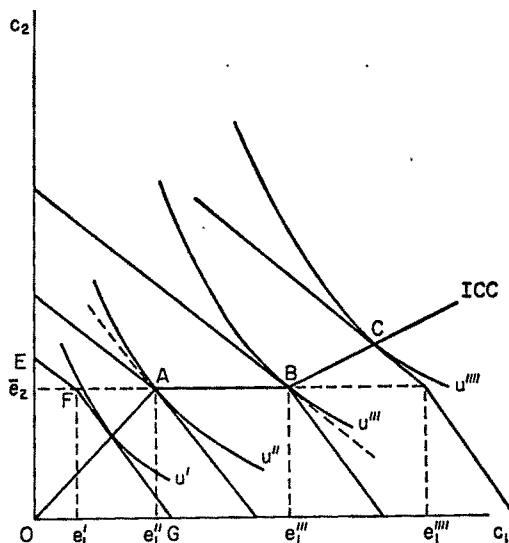


FIGURE 1

straight line from the origin.<sup>5</sup> The  $ICC$  in this figure is  $OABC$ . The line segment  $BC$  if extended leftwards would intersect the origin. The kinked lines like  $EFG$  are budget constraints (see Hirshleifer) with a slope of  $(-\rho_\alpha)$  above  $\bar{e}_2$  and a slope of  $(-\rho_\beta)$  below  $\bar{e}_2$ .

Along the line segment  $OA$  borrowing occurs. The consumer maximizes the expression:

$$(9) \quad u = u(e_1 + b, \bar{e}_2 - \rho_\beta b)$$

<sup>5</sup> That the income consumption curve is a ray from the origin may be shown easily. Using the theorem that any partial derivative of a function which is homogeneous of degree  $k$  is homogeneous of degree  $k-1$  we know that  $u_i(x_1, x_2)$ , ( $i=1, 2$ ) are homogeneous of degree zero. This may be written:

$$u_1(x_1, x_2) = u_1(ax_1, ax_2),$$

$$u_2(x_1, x_2) = u_2(ax_1, ax_2)$$

or,

$$\frac{u_1(x_1, x_2)}{u_2(x_1, x_2)} = \frac{u_1(ax_1, ax_2)}{u_2(ax_1, ax_2)}$$

Thus the marginal rate of substitution does not change along any ray from the origin. Since the  $MRS$  is equated to the price ratio (interest rate), in whatever ranges the price ratio remains fixed the  $ICC$  is a straight line from the origin.

This yields a solution of the borrowing function and real income function of:

$$(10) \quad b(e_1, \bar{e}_2, \rho_\beta) \geq 0$$

$$(11) \quad u_\beta^* = u^*(e_1, \bar{e}_2, \rho_\beta)$$

In any case where  $b(e_1, \bar{e}_2, \rho_\beta) < 0$ , the solution is not on the lower side of the budget constraint. Along the line segment  $BC$  he lends and maximizes:

$$(12) \quad u = u(e_1 + b, \bar{e}_2 - \rho_\alpha b)$$

and the borrowing function and real income function solve for:

$$(13) \quad b(e_1, \bar{e}_2, \rho_\alpha) \leq 0$$

$$(14) \quad u_\alpha^* = u^*(e_1, \bar{e}_2, \rho_\alpha)$$

Along the intermediate section,  $AB$ , he neither borrows nor lends. In this section the borrowing functions indicate the incompatible solutions  $b(e_1, \bar{e}_2, \rho_\beta) < 0$  and  $b(e_1, \bar{e}_2, \rho_\alpha) > 0$ .

The utility level reached along line segment  $AB$  is simply:

$$(15) \quad u_n^* = u(e_1, \bar{e}_2)$$

The income consumption curve represented here and the utility function provide the information necessary to find utility as a function of period one income holding the other factors constant. This will be expressed as:

$$(16) \quad u^{**}(e_1; \bar{e}_2, \rho_\beta, \rho_\alpha)$$

$$= \begin{cases} u^*(e_1, e_2, \rho_\beta) & \text{for } e_1 \leq e_1'' \\ u(e_1, \bar{e}_2) & \text{for } e_1'' < e_1 < e_1''' \\ u^*(e_1, \bar{e}_2, \rho_\alpha) & \text{for } e_1 \geq e_1''' \end{cases}$$

The function  $u^{**}(e_1; \bar{e}_2, \rho_\beta, \rho_\alpha)$  obeys the expected utility maximization axioms for period one income, and it is to this function that we shall devote our attention. This function is made up of three sections and two joining points. These will be examined individually, and the results are shown in Figure 2.

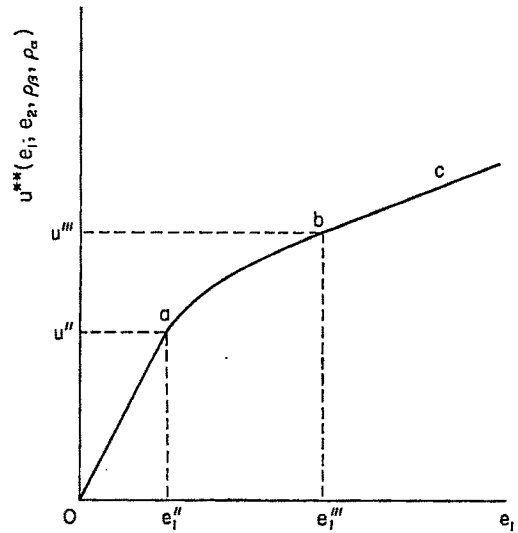


FIGURE 2

The first section,  $Oa$ , is for  $e_1 \leq e_1''$ . In this section the marginal utility of increasing income  $e_1$  is a positive constant. This is a consequence of moving out a linear homogeneous function along an isocline. Thus  $u^{**}(e_1; \bar{e}_2, \rho_\beta, \rho_\alpha)$  is a straight line with a positive slope for  $e_1 < e_1''$ .

The second section,  $ab$ , is for  $e_1'' < e_1 < e_1'''$ . In this section consumption in the first period is expanded while consumption in the second period remains constant. For a linear homogeneous utility function a diminishing marginal rate of substitution implies diminishing marginal utility.<sup>6</sup> This

<sup>6</sup> That linear homogeneity and a diminishing marginal rate of substitution,  $MRS$ , together imply diminishing marginal utility may easily be shown. The first step is to demonstrate that a diminishing  $MRS$  along an indifference curve implies a diminishing  $MRS$  for the expansion of a single factor letting utility vary if the utility function is linear homogeneous. That this should be true follows from fn. 5. If the  $MRS$  did not fall as the factor was increased, then the good must be an inferior good, but fn. 5 proves all iso-clines to be straight lines from the origin.

Start at the points  $(x_1, x_2)$  and  $(x_1', x_2')$  chosen such that  $u(x_1, x_2) = u(x_1', x_2')$ , and  $x_1' > x_1$ . Diminishing  $MRS$  implies that  $u_1(x_1, x_2)/u_2(x_1, x_2) > u_1(x_1', x_2')/u_2(x_1', x_2')$ . Take  $x_2'' = x_2$  and  $x_1''$  chosen such that  $x_1''/x_2'' = x_1'/x_2'$ . Then:

$$\frac{u_1(x_1, x_2)}{u_2(x_1, x_2)} > \frac{u_1(x_1', x_2')}{u_2(x_1', x_2')} = \frac{u_1(x_1'', x_2'')}{u_2(x_1'', x_2'')}$$

means that the function  $u^{**}(e; \bar{e}_2, \rho_\beta, \rho_\alpha)$  is upward sloping and concave for incomes between  $e_1''$  and  $e_1'''$ .

The final section,  $bc$ , is similar to the first section. Again the expansion is out an isocline. Thus  $u^{**}(e_1; \bar{e}_2, \rho_\beta, \rho_\alpha)$  is a positively sloped straight line in this region.

Since  $u^{**}(e_1; \bar{e}_2, \rho_\beta, \rho_\alpha)$  obeys the expected utility maximization postulates, three facts are known about this individual:

- that he will be risk-neutral for any gamble which involves only incomes that require borrowing (i.e.,  $e_1 < e_1''$  whether the gamble is lost or won);
- that he will be risk-averse for gambles after which he neither borrows nor lends; and
- that he will be risk-neutral if his income is high enough, win or lose in the gamble, for him to lend in either case.

There is one final step. This is to examine gambles of the following sort: Assume there is a .5 probability of getting  $(e_1 - \epsilon_1)$  and a .5 probability of getting  $(e_1 + \epsilon_1)$  where, for example,  $(e_1 - \epsilon_1)$  is in the neither borrowing nor lending zone and  $(e_1 + \epsilon_1)$  is in the lending zone. The examination must be for all pairwise choices of the borrowing zone, the lending

zone, and the neither borrowing nor lending zone.

The conclusions for these gambles are best arrived at by showing that  $u^{**}(e_1; \bar{e}_2, \rho_\beta, \rho_\alpha)$  is continuously differentiable. This may be demonstrated by showing that the left-hand derivative of  $u^{**}$  at point  $a$  is equal to the right-hand derivative at point  $a$  and that the left- and right-hand derivatives at point  $b$  are equal.<sup>7</sup>

This is simply demonstrated by referring back to Figure 1 where at point  $A$  the individual solves the problem:

$$(17) \quad \max_b \quad u(e_1'' + b, \bar{e}_2 - \rho_\beta b)$$

At point  $A$  the solution of this problem is  $b=0$  so:

$$(18) \quad u = u(e_1'', \bar{e}_2) \text{ at point } A$$

The first-order condition of the above maximization problem is:

$$(19) \quad u_1 - \rho_\beta u_2 = 0,$$

$$\text{where} \quad u_i = \frac{\delta u}{\delta c_i} \quad [i = 1, 2]$$

To find the value of the left-hand derivative at point  $A$  *totally differentiate*  $u(e_1'' + b, \bar{e}_2 - \rho_\beta b)$  assuming that  $\bar{e}_2$  is totally parametric:

$$(20) \quad \begin{aligned} du &= u_1 \frac{\delta c_1}{\delta e_1} de_1 + u_1 \frac{\delta c_1}{\delta b} db - \rho_\beta u_2 \frac{\delta c_2}{\delta b} db \\ du &= u_1 de_1 + u_1 db - \rho_\beta u_2 db \\ &= u_1 de_1 + (u_1 - \rho_\beta u_2) db \end{aligned}$$

But at the point  $(e_1'', \bar{e}_2)$ ,  $u_1 - \rho_\beta u_2 = 0$  so:

<sup>7</sup> The definition of a derivative of  $y=f(x)$  is:

$$\frac{dy}{dx} = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h}$$

A left-hand derivative will be defined for this limit as  $h$  approaches zero from the negative numbers; for a right-hand derivative,  $h$  will approach zero from the positive numbers.

Thus:

$$\left. \frac{d \left( \frac{u_1}{u_2} \right)}{dx_1} \right|_{du=0} < 0 \quad \text{implies that} \quad \frac{\delta \left( \frac{u_1}{u_2} \right)}{\delta x_1} < 0$$

The second step is to analyze this relationship:

$$\frac{\delta \left( \frac{u_1}{u_2} \right)}{\delta x_1} = \frac{u_{11}u_2 - u_{21}u_1}{u_2^2} < 0$$

For linear homogeneous functions  $u_{11}x_1 + u_{12}x_2 = 0$ . If  $u_{11} \geq 0$ , then the homogeneity condition would imply that  $u_{12} \leq 0$  whereas the diminishing MRS condition would imply that  $u_{12} > 0$ . We may reject the possibility of increasing marginal utility with these conditions.

$$(21) \quad \frac{du}{de_1} = u_1(e_1'', \bar{e}_2)$$

= (left-hand derivative at  $A$ )

To find the value of the right-hand derivative we need only find the partial derivative  $\partial u / \partial e_1$ . This is because  $b=0$  in this range and  $\bar{e}_2$  is parametric.

$$(22) \quad \frac{\delta u}{\delta e_1} = u_1(e_1'', \bar{e}_2)$$

= (right-hand derivative at  $A$ )

Thus the left-hand and right-hand derivatives of  $u^{**}$  are equal at the point  $A$ . Exactly the same proof may be supplied for the point  $B$  with the value  $e_1''$  used in the place of  $e_1'$  above.

Graphically this means that the von Neumann-Morgenstern utility function for gambling with present income is as shown in Figure 2.

The line segments  $Oa$ ,  $ab$ , and  $bc$  correspond to the line segments  $OA$ ,  $AB$ , and  $BC$ , respectively, in Figure 1. It is apparent that the intertemporally risk-neutral individual will be risk-averse for any gamble which involves more than one of these line segments. He acts as if he were risk-averse for many present gambles.

For a more complex imperfect capital market I shall present the results graphically, without proof. Two assumptions are retained: that  $\bar{e}_2$  is known and fixed, and that the consumption decisions are made after the results of all gambles are known. Suppose the capital market is such that larger borrowing entails higher interest rates<sup>8</sup> and that larger amounts invested yield at first increasing rates of return and eventually decreasing rates of return.<sup>9</sup> The

<sup>8</sup> This may be in the form of "compensating balances" required by lending institutions or in the form of credit rationing, e.g., an infinite interest rate (see Dwight Jaffee and Franco Modigliani, and *The Wall Street Journal*).

<sup>9</sup> Assume that the first dollars may be invested in a bank savings account. As costs of investment decrease

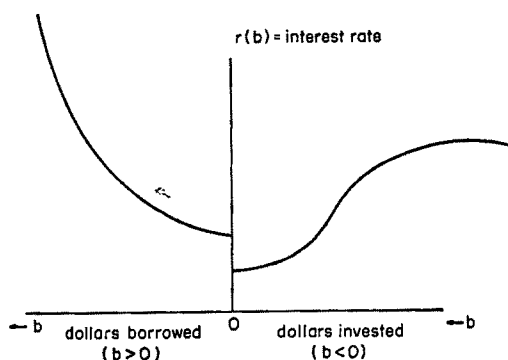


FIGURE 3

capital market may be characterized by the following structure of interest rates,  $r(b)$ , shown in Figure 3.

In this case the von Neumann-Morgenstern utility function of an individual who is risk-neutral for all-period consumption, as defined above, will have a form for present income very similar to the Friedman-Savage form (see Friedman and Savage, pp. 57-96.)<sup>10</sup>

With these two conditions I assert, without proof, that the utility function for present income looks like Figure 4. (A specific case from the risk-loving portion of this curve is shown in the Appendix.)

In such a case a person will be closer to his lower inflection point on the left as he becomes less of a net creditor. The reader may note that, in general, the individual stays near his inflection point, as shown in Harry Markowitz's article. This conclusion is strengthened as transactions costs for borrowing and lending are introduced.

Finally, an assumption may be made

as a proportion of total investment, money will be invested in stocks or productive assets. The eventually declining portion of the lending curve is the familiar downward sloping marginal efficiency of investment curve. The eventually decreasing interest rate is not needed for the risk loving to exist. It corresponds to the eventual reestablishment of a high income (not wealth) risk-averse section.

<sup>10</sup> This requires the assumption of convex indifference curves. Since, if indifference curves were not convex, individuals might consume their whole wealth in only one time period, this is a weak assumption.

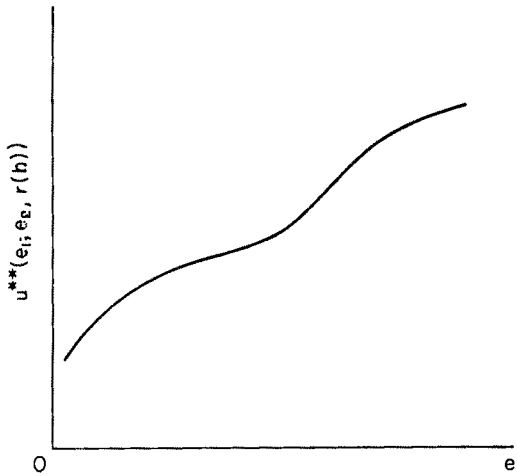


FIGURE 4

that enables us to predict behavior in addition to that predicted by the Friedman-Savage model. This assumption is that  $r(b)$  for borrowing is primarily a function of future income from physical assets. This assumption is just that lending firms wish borrowers to show evidence of a high probability of repayment. Higher future incomes are made up of return from physical assets, return from human capital, and cash value of assets sold (e.g., at death.) The lower a person's asset level, the higher interest he must pay for any given absolute level of money borrowed.

The argument presented here is thus that, *ceteris paribus*, net debtors should behave in a more risk-averse fashion than net creditors. In fact, risk aversion should be more pronounced as net debt increases. This statement must of course be modified in countries with bankruptcy laws. Where declaration of bankruptcy is legal, there is a point at the left of this transformed utility function for present income where the function becomes flat, i.e., the individual becomes a risk lover. In countries without bankruptcy laws the utility function becomes flat at the lower bound given by survival. Since the first derivative of the utility function at this point *may* not exist (e.g., a kink with a flat portion to the left),

risk taking *may* only be observed for the ultimate desperate move. This case I shall not consider further although we all have read of some mild hero succeeding (or not) in one final attempt.

## II. Using the Model

The model as presented relates the risk-neutral individual's utility to present period income if he faces the given structure of interest rates. Other factors which may amplify the basic results of this model are transactions costs in the capital markets and transactions costs in the markets for physical capital. The existence of transactions costs may be partly reinterpreted above as a bigger displacement between the borrowing and lending rate of interest. Thus we would expect the higher proportion of illiquid capital a firm has, the more risk-averse its entrepreneur will be.

C. M. Elliot, in his work on African development, advances a hypothesis similar to the one above. He asserts that a primary reason for the more successful cropping of cotton in Uganda than in Kenya was due to institutionally determined risk behavior, which in Uganda did a more efficient job of spreading the risk. In Kenya the individual had no good source of borrowing in the case of failure. In Uganda the tribesmen could borrow from the headman in the case of a bad year. Elliot feels that Kenya might have been similarly successful in cotton growing had a similar capital market been available.

This model may also yield some insight into the reputation of Chilean peasants who are said to be highly conservative in cropping patterns but willing to gamble with their earnings.

## III. Toward Less Risk Aversion and Greater Progressivity

One policy question of interest is how a country may promote more risk-neutral behavior on the part of its lower-asset, lower-income population. This is the prob-

lem of promoting more progressivity of the subsistence farmer or of the ghetto entrepreneur. These are important policy problems for the underdeveloped country and for the urban poor of our own country.

The policy prescriptions of this model for promoting intra-sectoral growth follow simply from the factors contributing to risk aversion. The first of these is to create more nearly perfect borrowing markets. In many instances this means a locational shift in the banking system to have a higher density near the farm areas or the ghetto areas. It may also mean a reduction in red tape surrounding loan procedure. In many cases, peasants or individuals who cannot read nor write may not feel safe borrowing from a bank under present conditions. Finally, the central government may have to insure or cosign loans for the individual farmers. Cosigners may be a scarce commodity for the low-level farmer or businessman.

The second policy prescription is to protect the individual with a welfare program. The person who loses everything if he errs is likely to be more risk-averse than the individual who goes on the welfare roles if he errs. And the person on or close to welfare will (in some income ranges) be more risk-averse than one under a system of negative income taxes. These policies must also be supplemented by a form of limited liability and bankruptcy laws. If welfare payments or if any income after leaving welfare roles is in jeopardy, then the individual must be more risk-averse.

Finally there is the use of subsidization. In the United States, government contracts are now subsidizing black entrepreneurship. Similarly, countries often subsidize agricultural innovation by decreasing the cost of inputs. Another common type of subsidy is price supports. If the government insures that it will buy a crop's output at some minimum base price, then a farmer's expected value of return

from the crop is increased and the variance of the return is decreased. The government, if it properly selects innovations for this type of subsidization, will have low expected costs. The farmer's expected value of any innovation the government wishes to subsidize will be higher than the minimum payments promised to the farmer, and thus the payments for failure should, on the average, be very far below the promises to pay in the case of failure. *Of course care must be taken in such a program not to set the minimum payment too high.* If the minimum payment is above a certain level, then there will be less incentive to nurture properly the new crop.

In an underdeveloped country increased agricultural production may be a precondition for sustained growth in all sectors. In our country many people feel that growth of black entrepreneurship should be encouraged for moral, political, and economic reasons. If greater progressivity can flow from reducing risk-averse behavior (i.e., to the level of risk-neutral behavior), then the rationale and tools presented here should be used to a greater extent in economic policy.

#### APPENDIX

##### *An Example of a Gamble for an Intertemporally Risk-Neutral Individual*

I shall heuristically present the outcome of a single gambling situation. For ease of graphical construction this gamble will have an expected value at the level of income at which the individual is indifferent between borrowing his first dollar or not borrowing at all.

For simplicity of analysis the structure of interest rates will be assumed to look like Figure 5. The more general case depends upon the actual curvature of the left-hand portion of this curve, but the results here are generalizable to the case presented in Figure 3 of the text.

If the individual has endowment income in period 1 of  $\bar{e}_1$  and endowment income in

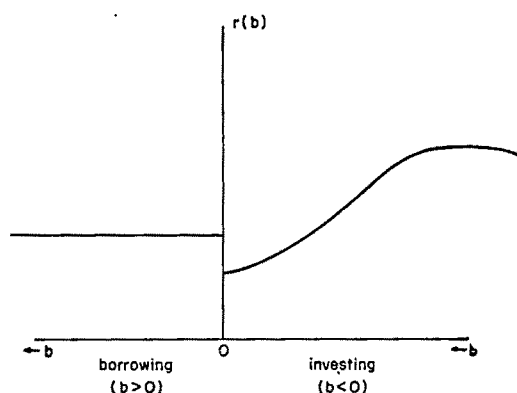


FIGURE 5

period 2 of  $\bar{e}_2$ , then his budget constraint has the shape shown in Figure 6; where the slope of the budget line is  $-(1+r(b))$ , and  $(+b)$  is the horizontal distance to the right of  $\bar{e}_1$  to the constraint.

Now one gamble may be presented (see Figure 7). In this case the individual is given a 50-50 gamble between  $\bar{e}_1 - \epsilon$  and  $\bar{e}_1 + \epsilon$  and prefers the gamble to  $\bar{e}_1$  with certainty. This is because the utility function is linear homogeneous. The linear homogeneity implies that the indifference curves  $a$ ,  $b$ , and  $c$  are equidistant in utility terms. In other words a person offered 1) a 50-50 gamble be-

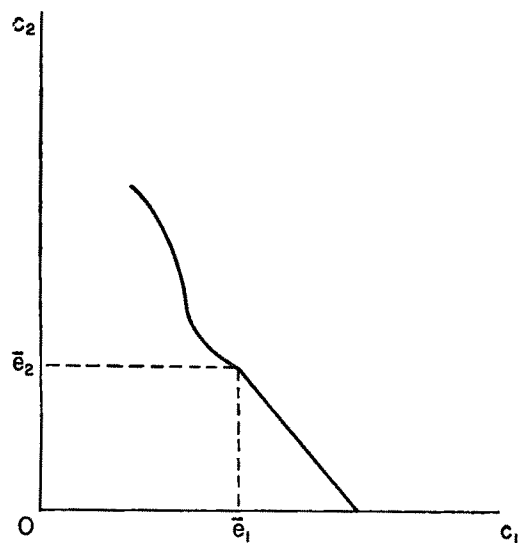


FIGURE 6

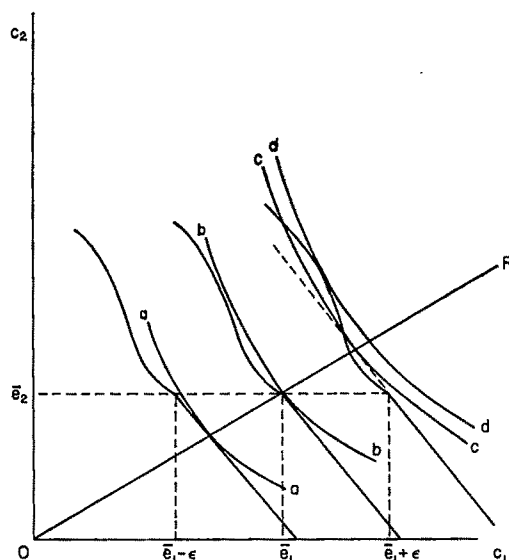


FIGURE 7

tween utility levels  $a$  and  $c$  or 2) the utility level  $b$  with certainty would be indifferent between the alternatives. Since the person given the gamble between  $\bar{e}_1 - \epsilon$  and  $\bar{e}_1 + \epsilon$  is being offered an even odds gamble between the utility levels  $a$  and  $d$ , and the level  $d$  is greater than the level  $c$ , he will pick the gamble in preference to level  $b$  with certainty.

This establishes the existence of risk loving for current income given an intertemporally risk-neutral individual.

# REFERENCES

- W. J. Baumol, *Economic Theory and Operations Analysis*, Englewood Cliffs 1965.
- J. M. Boussard and M. Petit, "Representation of Farmers' Behavior with a Focus-Loss Constraint," *J. Farm Econ.*, Nov. 1967, 49, 869-81.
- C. M. Elliot, "Agricultural and Economic Development in Africa: Theory and Experience, 1880-1914," in E. L. Jones and S. J. Woolf, eds., *Agrarian Change and Economic Development*, London 1969.
- M. S. Feldstein, "Mean-Variance Analysis in the Theory of Liquidity Preference and Portfolio Selection," *Rev. Econ. Stud.*, Jan. 1969, 36, 5-14.

- I. N. Fisher and G. R. Hall, "Risk and Corporate Rates of Return," *Quart. J. Econ.*, Feb. 1969, 83, 79-92.
- M. Friedman and L. J. Savage, "The Utility Analysis of Choices Involving Risk," *J. Polit. Econ.*, Aug. 1948, 56, 279-304.
- N. Hakansson, "Friedman-Savage Utility Functions Consistent with Risk Aversion," *Quart. J. Econ.*, Aug. 1970, 84, 472-87.
- J. Hirshleifer, "On the Theory of Optimal Investment Decisions," *J. Polit. Econ.*, Aug. 1958, 66, 329-52.
- D. M. Jaffee and F. Modigliani, "A Theory and Test of Credit Rationing," *Amer. Econ. Rev.*, Dec. 1969, 59, 784-93.
- S. R. Johnson, "A Re-examination of the Farm Diversification Problem," *J. Farm. Econ.*, Aug. 1967, 49, 610-21.
- H. Markowitz, "The Utility of Wealth," *J. Polit. Econ.*, Apr. 1952, 63, 151-59.
- D. H. Pyle and S. J. Turnovsky, "Safety-First and Expected Utility Maximization in Mean-Standard Deviation Portfolio Analysis," *Rev. Econ. Statist.*, Feb. 1970, 52, 75-81.
- G. J. Stigler, "Imperfections of the Capital Market," in G. Stigler, ed., *The Organization of Industry*, Homewood 1968.
- J. E. Stiglitz, "Behavior Towards Risk with Many Commodities," *Econometrica*, Oct. 1969, 37, 660-67.
- B. P. Stigum, "Entrepreneurial Choice Over Time Under Conditions of Uncertainty," *Int. Econ. Rev.*, Oct. 1969, 10, 426-42.
- L. Telser, "Safety-First and Hedging," *Rev. Econ. Stud.*, 1955-56, no. 1, 23, 1-16.
- J. Tobin, "Liquidity Preference as Behavior Towards Risk," *Rev. Econ. Stud.*, Feb. 1958, 25, 65-86.
- The Wall Street Journal*, Mar. 9, 1970, p. 28.

# Production, Trade, and Protection When There are Many Commodities and Two Factors

By WILLIAM P. TRAVIS\*

When production functions are homogeneous of the first degree, products are numerous, and only two factors of production exist, the production possibilities surface (or, for brevity, simply the production surface) of the economy has a special and simple shape which is easily described by means of a two-dimensional diagram. This representation clarifies the relationships among specialization and relative prices, on the one hand, and the indeterminacy of the international localization of production on the other. Indeterminacy in turn contributes to the understanding of the dynamics of commercial policy.

The possibility of indeterminacy was first raised by Paul Samuelson (pp. 8-10 and 12-14). I have discussed some of its commercial-policy implications elsewhere (ch. 4); and they have recently been amplified, for the three-output case, by James Melvin. His extremely lucid analysis of this case and emphasis of its importance have inspired the present analysis of the case of two factors and many commodities.

Though it is still special from a general theoretical viewpoint, this case is realistic. Capital (including human capital, which raises salaries above wages) and labor each earn about one half of the income of just about any country, while the contribution of all types of land is only a few percent. A multitude of products clearly exists, meanwhile, and is recognized, notably, by

the list itself of national tariff rates. Thus the special case of very many commodities and only two (important) factors is essential for protection theory. The somewhat odd behavior it implies (see Murray Kemp, p. 50) perhaps should be given more attention.

The description of the multiproduct production surface requires only the construction of a series of two-commodity production curves from the corresponding ordinary, two-product contract curves. This means that one can construct the national production surface from available data and thus analyze the full general-equilibrium allocative and price effects of tariffs. In particular, it is possible to measure the general restrictiveness of the tariff. We shall see for example that the U.S. tariff prohibits the exchange of capital-intensive for labor-intensive products.

## I. The Production Surface

The derivation of the  $n$ -dimensional production surface is based on the following model of production:

$$(1) \quad 1 = A_i \frac{L_i}{x_i} \psi_i(\rho_i), \quad (i = 1, \dots, n)$$

$$\sum_{i=1}^n L_i = L = E_1,$$

and

$$\sum_{i=1}^n L_i \rho_i = K = E_2$$

where  $A_i$  is a scalar indicating the neutral efficiency of the  $i$ th production function,

\* University of California at San Diego. I am indebted to John Cambron, referees, and the editor for many improvements.

$L_i$  the amount of labor and  $L_i\rho_i$  the amount of capital which that function employs,  $\rho_i$  the capital-labor ratio of the  $i$ th industry or product,  $x_i$  the amount of product  $i$  produced,  $L$  the total labor employment,  $E_1$  the total labor supply,  $K$  the total capital employment, and  $E_2$  the total capital supply. Equations (1) are homogeneous of the first degree, which permits them to be cast in terms of the unit outputs to give the expressions for the unit isoquants of the production functions.

Using these isoquants, we begin by constructing each of the two-by-two production curves of the economy. Each of these *conditional production curves* is drawn on the assumption that the outputs of the remaining  $(n-2)$  products are all held to zero and thus it lies in its own two-dimensional axial plane. Each curve corresponds point-for-point (see Kurt Savosnick) under the assumptions made above with a contract curve traversing the Edgeworth box of dimensions  $(E_1, E_2)$ . There are  $\frac{1}{2}n(n-1)$  (the number of separate pairs of any  $n$  objects) such contract curves and an equal number of corresponding conditional production curves.

Figure 1 illustrates the derivation of the conditional production curve for product 1 and product  $n$  (products are numbered in order of their labor intensiveness,  $1/\rho_i$ ) from their corresponding unit isoquants, labelled  $i_1$  and  $i_n$ , respectively, and the national endowment vector  $E = (E_1, E_2)$ .

The first step is to draw the contract curve  $(n, 1)$  which, in Figure 1, the curve  $Oa_{1n}E$  represents. Pick an arbitrary relative wage rate  $\omega = w/r$ , such as that indicated by the slope of the line  $X_nEX_1$ . The input ray  $OX_n$  is then chosen to intersect the unit isoquant  $i_n$  in the point,  $n$ , where its slope is parallel to  $X_nEX_1$ . Similarly, the input ray  $OX_1$  is chosen to intersect  $i_1$  in the point 1, where it has that common slope. The intersection of the ray  $Ea_{1n}$  (which is parallel to  $OX_1$ ) with ray  $OX_n$

thus lies in the contract, or efficiency, locus, and is marked  $a_{1n}$ . By rotating the factoral budget line containing  $E$  about  $E$  and repeating the above procedure, one traces out the entire contract curve, each point of which therefore corresponds to a particular relative wage rate.

The point  $a_{1n}$  indicates a definite output of each commodity. The output of the  $n$ th commodity is equal to the ratio

$$\frac{Oa_{1n}}{On},$$

inasmuch as the point  $n$  lies in the unit isoquant. The output of the first commodity is equal to the ratio

$$\frac{Ob}{OI}$$

inasmuch as the point 1 on the input ray  $OIbX_1$  lies in the unit isoquant of the first commodity, and the vector  $Ob$  is equal to the vector  $Ea_{1n}$ . Thus a point in output space corresponds to the point  $a_{1n}$  on the contract curve in input space. Call the output point  $q_{1n}$ ; it lies on the conditional production curve, and that curve is traced out by repeating the above procedure for every point in the contract curve. Observe that each point  $a_{1n}$  and thus  $q_{1n}$  corresponds to a given relative wage rate, indicated by the slope of the factoral budget line  $X_nEX_1$  or

$$wE_1 + rE_2 = Y = wL + rK = rL(\omega + \rho)$$

containing the point  $E = (E_1, E_2)$ .

The reader can now imagine all  $n$  of the unit isoquants drawn onto Figure 1, together with the corresponding  $\frac{1}{2}n(n-1)$  conditional contract curves. These contract curves are the factoral-space, point-for-point, images of the intersections of the production surface with the axial planes of commodity space. Because of the "flatness" of the production surface, they suffice to describe it.

The way in which they do so is illus-

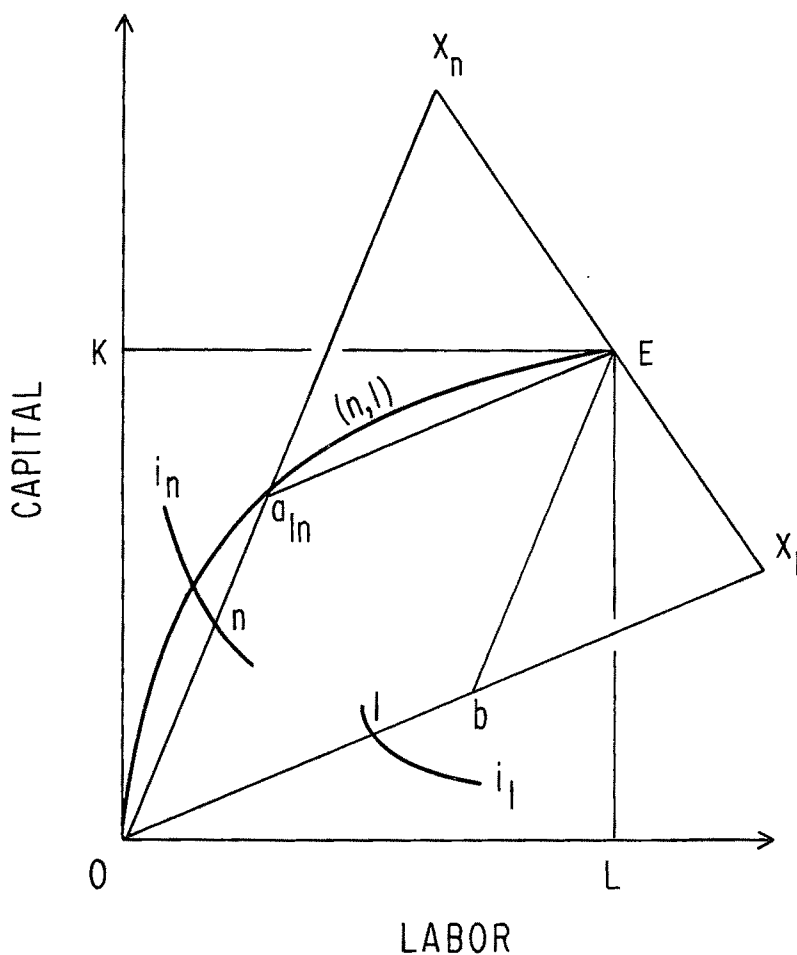


FIGURE 1

trated by Figures 2 through 4. Figures 2 and 3, following Melvin, provide the illustration for the case of three commodities and two factors, which can be visualized directly. The three contract curves, in Figure 2, are labelled according to their pairs of commodities: (3, 1), (2, 1), and (3, 2). The three corresponding conditional production curves are drawn in perspective in Figure 3 as  $Q_3Q_1$ ,  $Q_2Q_1$ , and  $Q_3Q_2$ , respectively.

When the relative wage rate is that indicated by the slope of the line  $X_3X_2EX_1$  of Figure 2, the input rays  $OX_3$  and  $OX_2$  both lie to the left of the endowment

point  $E$ , and the endowment ray  $OX_1$  to its right. The ray  $Ea_{13}$ , parallel to ray  $OX_1$ , intersects the other two rays in the points  $a_{13}$  and  $a_{12}$ , which lie on the contract curves (3, 1) and (2, 1), respectively. The contract curve (3, 2) begins to be traced out by such intersection points only when the relative wage rate is low enough so that the rays  $OX_3$  and  $OX_2$  lie on opposite sides of  $E$ , as the various straight lines rotate, and their intersections move, in the senses which the small arrows describe.

The unit isoquants (not shown) can be used now in the same manner as that described above to obtain from the intersec-

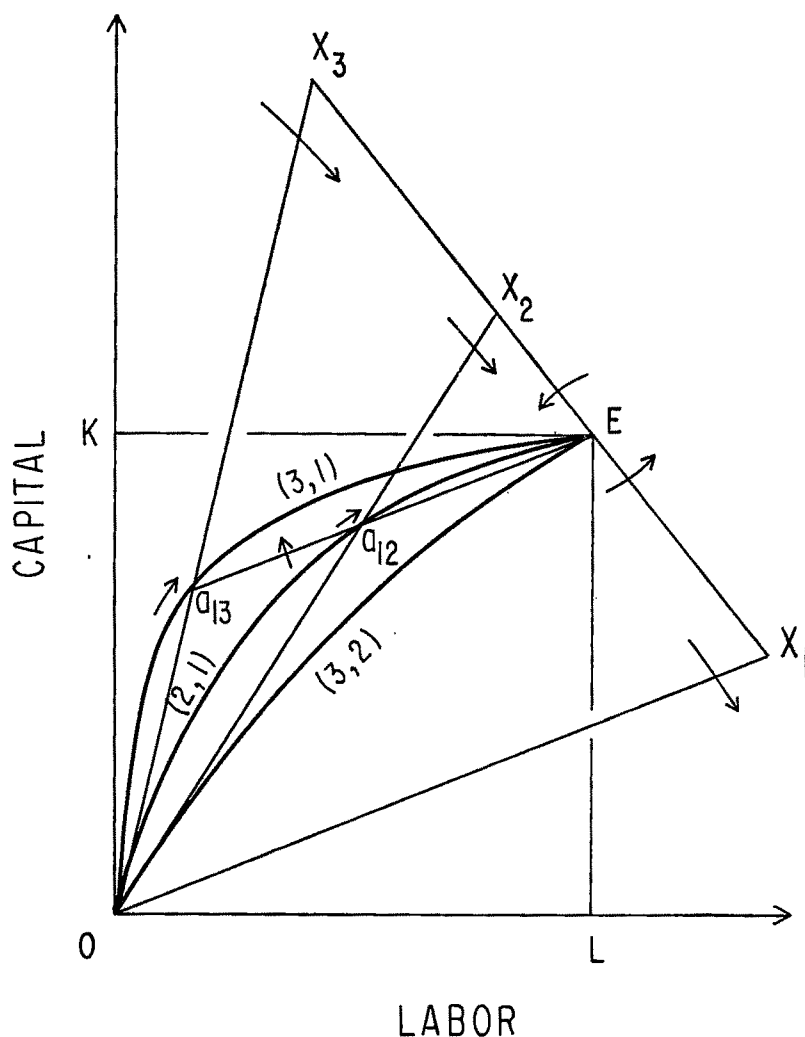


FIGURE 2

tion points,  $a_{hj}$ , the corresponding output points,  $q_{hj}$ , in commodity space, as Figure 3 illustrates. Thus the coordinates of point  $q_{13}$  in Figure 3 are  $a_3$  and  $a_{13}$ , in the third and first commodity axes. The point  $a_3$  lies the same number of units from  $O$  on  $OQ_3X_3$  as the point  $a_{13}$  lies from  $O$  along the ray  $OX_3$  in Figure 2, where units along  $OX_3$  are defined in terms of the distance from point  $O$  to the intersection of  $OX_3$  with the unit isoquant of the third commodity. The point  $a_{13}$  in Figure 3 lies the same number of units from  $O$  on the axis

$OQ_1X_1$  as the point  $a_{13}$  in Figure 2 lies from  $E$ , where distance along ray  $Ea_{13}$  is measured in terms of the distance along that ray between the integer isoquants of the first product.

The point  $q_{12}$  in Figure 3 corresponds in the same manner to the intersection  $a_{12}$  in Figure 2. The first coordinate of  $q_{12}$  is the point  $a_{12}$ , and the segment  $a_{13}a_{12}$  of the vertical axis is thus the direct image of the segment  $a_{13}a_{12}$  of the input ray,  $Ea_{12}a_{13}$ , of the first commodity in Figure 2. Corresponding to this segment is the line seg-

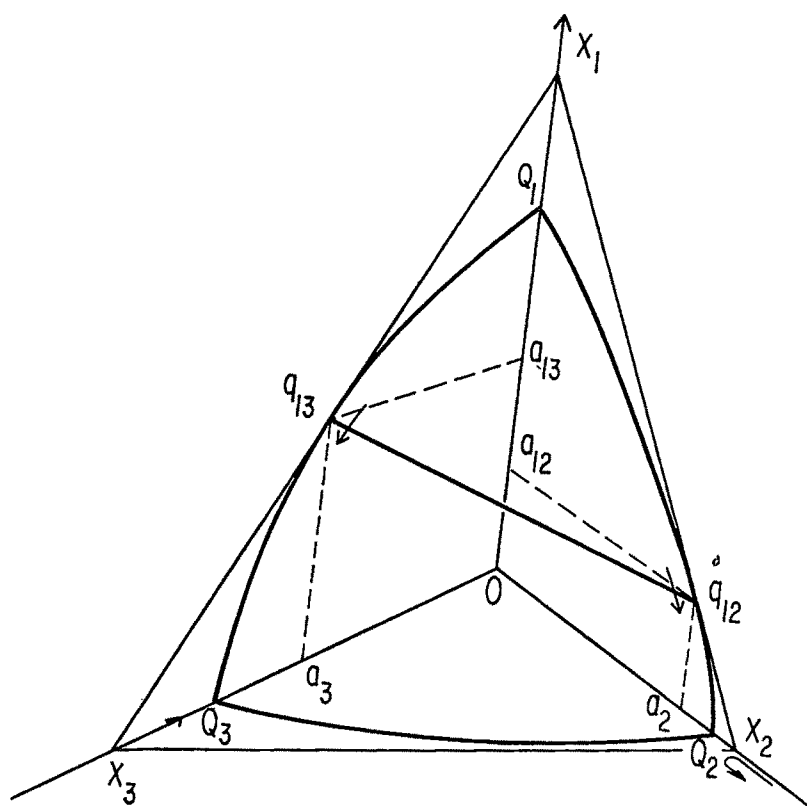


FIGURE 3

ment  $q_{12}q_{13}$  in Figure 3, which lies in the production surface  $Q_1Q_2Q_3$ . This is necessarily a straight line segment inasmuch as the same small change in the output of the first commodity will always result in the same change in the outputs of the second and third commodities, no matter where one starts on the segment  $a_{12}a_{13}$  (Figure 2), which is just the definition of a straight line.

We now have enough information to trace out the entire production surface as the relative wage is allowed to vary within the permissible range. The permissible range lies between the slope of the isoquant map of the first product, where its highest isoquant contains the point  $E$ , and the slope of the highest isoquant of the last (third) product where it contains  $E$ . For each permissible relative wage rate, such as that which Figure 2 indicates, we

obtain directly two points given by the intersection of the production rays emanating from the point  $E$  with the three contract curves. In the three-commodity case, there can never be more than two such intersections and they give the end-points of the output line segment joining the points on the two appropriate conditional production curves in commodity space. The arrows in Figure 3 help to visualize the process of tracing out the production surface as the relative wage rate alters within the permissible limits.

The output lines such as  $q_{12}q_{13}$  in Figure 3, being straight, must lie in the budget surfaces:

$$Y = p_1q_1 + p_2q_2 + p_3q_3 = \sum_{i=1}^3 p_iq_i$$

which are tangent to the production sur-

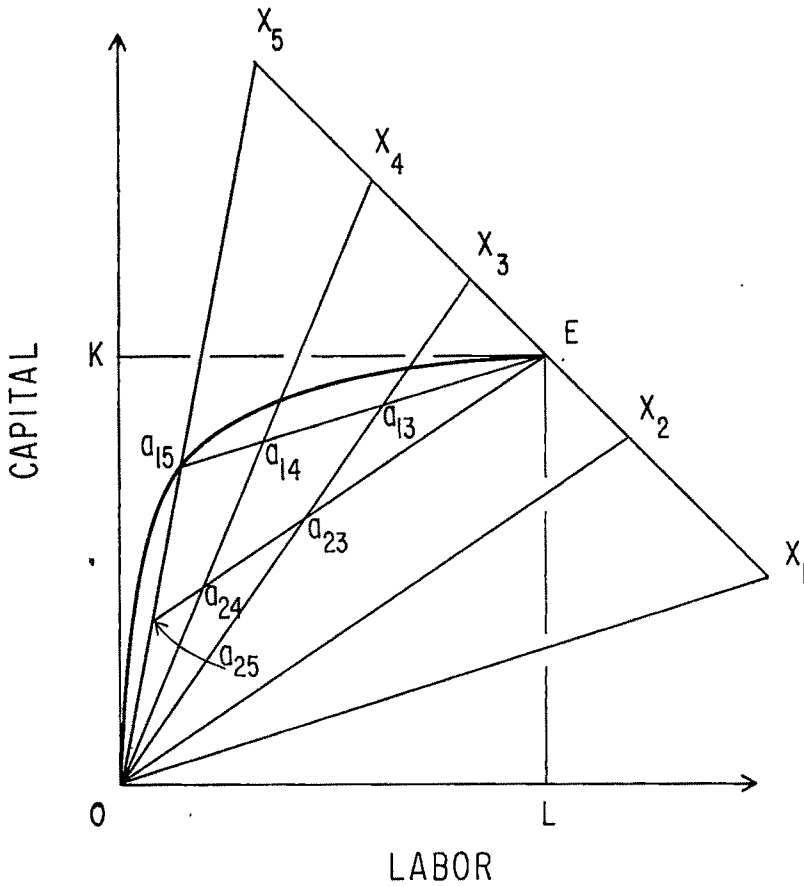


FIGURE 4

face. Before tackling the  $n$ -commodity case, it will be instructive to examine the relationship between these tangential budget planes and the production surface. The outputs corresponding to the input points  $X_1$ ,  $X_2$ , and  $X_3$  in Figure 2 are those which use quanta of the factors worth just the factoral budget,  $Y$ , at the wage and interest rate pertaining to the budget line through the point  $E$ . The corresponding outputs, because Euler's theorem applies, are consequently worth  $Y$  dollars and are shown in Figure 3 as the points  $X_1$ ,  $X_2$ , and  $X_3$ , respectively, lying in the commodity axes. Because each of those amounts is worth  $\$Y$ , the points determine the commodity budget plane  $X_1X_2X_3$ . Because the output point  $q_{13}$  contains quantities of the

first and third commodities which also use  $Y$  dollars' worth of resources, this output point, as well as point  $q_{12}$ , must also lie in the  $\$Y$  commodity budget plane, and so must all output points in the straight line segment  $q_{12}q_{13}$ .

A budget plane which contains at least two points in common with the production surface thus contains a whole line segment of common points. The prices,  $p_i$ , determining the partial slopes of such a budget plane can be called *consistent* with the factoral wages,  $w$  and  $r$ , determining the slope of the factoral budget line containing the point  $E$ . Price-wage consistency is important in the study of protection, as we shall see presently.

The determination of consistent prices

is direct. Because the output,  $X_i$ , of the  $i$ th commodity corresponding to the input point  $X_i$  determined by the intersection of the  $i$ th input ray with the factoral budget line absorbs  $\$Y$  of inputs, it is worth  $\$Y$ . Consequently  $p_i = Y/X_i$ .

The output  $X_i$  can be read off from the factoral space diagram directly by using the successive unit isoquants of the  $i$ th product to graduate the input ray  $OX_i$ . An evocative way of visualizing the relationship between the factoral budget line and the consistent commodity budget plane is therefore to think of the input rays in factoral space as if they were the corresponding axes in output space, graduated by their intersections with the isoquants of integer amounts of the commodities. The perspective of this factor-space image of output space is then such that the commodity budget plane is viewed exactly edgewise and projected into the straight factoral budget line.

An examination of Figures 2 and 3 reveals a further relationship which will aid in the  $n$ -commodity generalization. Observe that all consistent budget planes must be tangent to the conditional production curve (3, 1), or  $(n, 1)$ . So long as the relative wage rate is such that the input ray  $OX_2$  lies to the left of the endowment point, all the consistent budget planes also lie tangent to the conditional production curve (2, 1), as can be seen by the intersection of the relevant input rays in the contract curve (2, 1). When the input ray  $OX_2$  falls to the right of  $OE$ , output points corresponding to the contract curve (2, 1) will no longer be worth the factoral budget, and the consistent budget plane must be tangent to the conditional production curve (3, 2). It can be seen that the reason that the (3, 1) contract and conditional production curves are always involved is that the corresponding input rays straddle the endowment ray for all relative wage rates that do not enjoin complete specialization in either of the extreme products 1

or  $n$  (in which case one or the other corresponding input ray is the endowment ray). In the  $n$ -commodity case, therefore, the  $(n, 1)$  conditional production curve always supports the consistent budget plane.

This consideration facilitates the statement of conditions governing relative prices and the relative wage, for one has only to revert then to the basic diagram of Figure 1. The  $(n, 1)$  conditional production curve will accordingly be referred to as the principle conditional production curve.

The above methods of constructing and visualizing the three-dimensional production surface, although perhaps not the most efficient for this limited case, generalize easily into any number of dimensions and thus provide the means for analyzing more comprehensive cases involving indeterminacy. Obviously one cannot draw in all  $\frac{1}{2}n(n-1)$  contract curves and the accompanying web of input rays when  $n$  is large. This is unnecessary, however, because of the key role of the principle contract curve and the easy manner of obtaining all consistent prices from the factoral budget line.

These prices then give directly the equation for the consistent commodity budget hyperplane:

$$\$Y = \sum_{i=1}^n p_i x_i = p'x$$

From the contemplation of the three-dimensional case, we know that some of the points in the tangent hyperplane  $Y = p'x$  lie in the production hypersurface,  $T(q) = 0$ . For most practical problems, it is enough to know what is the intersection of the sets  $\{q | T(q) = 0\}$  and  $\{x | p'x = Y\}$ . For this set the ultra-simple formula of the consistent budget hyperplane directly provides the formula for the relevant portion of the mathematically more complicated production surface.

Because the tangency lines of the budget hyperplane and the production hypersurface extend to the principle conditional production curve and to certain of the other conditional production curves, one can border the region of their tangency. The other conditional production curves upon which the consistent budget hyperplanes rest are given, meanwhile, by the contract curves and the production rays.

Figure 4, which arbitrarily takes  $n=5$ , illustrates this determination. The slope of the factorial budget line gives, by its tangencies with the isoquant maps, the five input rays,  $OX_1$  through  $OX_5$ . The principle conditional production curve is accordingly  $(5, 1)$ , of which the contract curve is shown in Figure 4. Because the endowment ray  $OE$  in the illustration separates the first two production rays from the remaining three, a total of  $2 \times 3 = 6$  of the  $\frac{1}{2}(n-1) = 10$  contract and conditional production curves will be relevant for the depicted relative wage rate. The six input points  $a_{hi}$  ( $h=1$  and  $2$ ;  $i=3, 4, 5$ ) in Figure 4 mark the six extreme, alternative input points consistent with the condition that the input vector  $OE$  be fully employed at the input proportions corresponding to the input rays. The six corresponding points  $q_{hi}$  in output space thus lie both in  $T(q)=0$ , the production surface, and in  $p'x=Y$ , the consistent budget hyperplane.

Because there are six common points but  $n$  is only 5 we observe the possibility, absent in the special case of  $n=3$ , that the budget hyperplane may be tangent to the production hypersurface in a *region* (rather than merely a subspace) of the budget hyperplane (and common to the production surface) having the same dimensionality,  $(n-1)$ , as the budget hyperplane. If this is true, the importance of indeterminacy is increased.

To check out this possibility, it is necessary to ensure that the six vectors  $q_{hi}$  are

distinct, in which case any five of them span five-space, and all six must lie in the four-dimensional subspace common to the budget hyperplane and the production surface.

To check whether the six output vectors form a basis for five-space, observe first that each of them lies in an axial plane, then form the following table:

$$\begin{Bmatrix} q_{13} & q_{14} & q_{15} \\ q_{23} & q_{24} & q_{25} \end{Bmatrix}$$

and, finally, observe that five of the vectors, as indicated by the underlined subscripts, can be associated each with its own axis. Thus those five vectors form a basis of five-space, and the production surface coincides with the budget plane in a four-dimensional region of the latter, and not merely with a lower dimensional subspace of it. One can easily verify that this property holds also for the case of  $n=4$ , provided that the endowment ray separates two of the input rays from the rest.

Whatever the value of  $n$ , however, the relative wage rate  $\omega$  can be so extreme (high or low) that the tangency of the budget hyperplane with the production hypersurface is a subspace of the budget hyperplane. Obviously this is true if  $\omega$  is, say, so high that the input ray of the most labor intensive product, product 1, coincides with the economy's endowment vector,  $OE$ , in which case only the first product is produced and the tangency of the budget hyperplane and the production surface is a single point.

When, furthermore,  $\omega$  separates either the first or the  $n$ th input ray from the rest, the dimensionality of the tangency region is always  $(n-2)$ , regardless of  $n$ . Figure 5 illustrates this proposition for the case  $n=5$  and  $\omega$  so high that only the first input ray lies to the right of the endowment ray. In this case the ray  $Ea_{15}$ , parallel to the

first input ray, intersects the four remaining input rays, and thereby establishes four output vectors  $q_{1i}$  lying in both the budget hyperplane and the production surface. It is easy to see that in general  $(n-1)$  such intersections exist when either extreme input ray is separated from the rest. One can associate only  $(n-2)$  distinct output points with the  $(n-1)$  intersections, however, because one intersection must be associated with the first commodity axis. Because the budget hyperplane has dimensionality  $(n-1)$ , its tangency with the production surface is a hyperplane of dimensionality  $(n-2)$ .

Whatever the values of  $\omega$  and  $n$ , the boundary lines of the region or subspace of tangency belong to the set of lines determined by taking the convex combinations of the output points,  $q_{hi}$ , corresponding to the intersectional input points  $a_{hi}$ :

$$dq_{hi} + (1-d)q_{jk} = y$$

$$0 \leq d \leq 1$$

and

$$k \neq i \quad \text{when } h = j;$$

$$j \neq h \quad \text{when } k = i$$

There are  $m(n-m)$  intersectional points  $a_{hi}$ , when the endowment ray separates  $m$  of the  $n$  input rays from the rest, and thus  $\frac{1}{2}(m^4 - 2m^3n - m^2n^2 - mn^3 + m^2)$  lines  $dq_{hi} + (1-d)q_{jk} = y$ . Any such line is easy to determine from the two appropriate contract curves and sets of input rays, and thus one can easily explore any desired area of the region of tangency of the production hypersurface and the budget hyperplane.

In particular it can be seen, in answer to a dubiety which Melvin expressed (pp. 1265 and 1267), that indeterminacy of the production structure under a given relative wage rate persists into higher dimensions and continues to be important. In fact, one can consolidate the various points explored above to state the following proposition: There is a budget hyperplane roughly

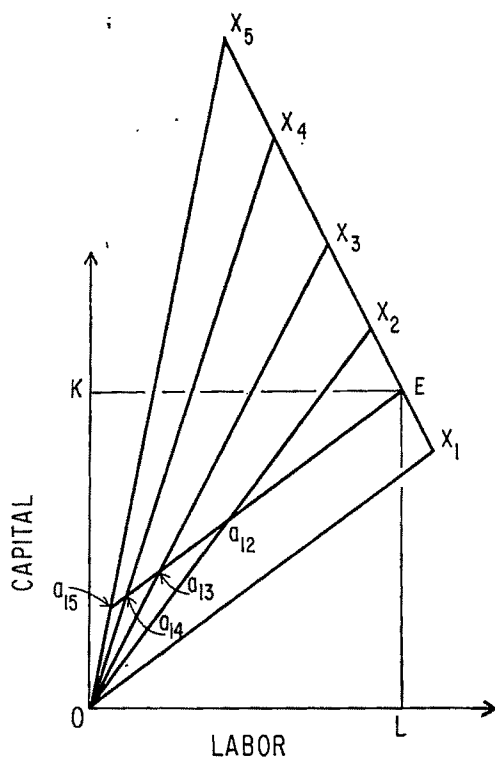


FIGURE 5

half of which coincides, when  $n$  is large, with the production surface.

The proof of this proposition is as follows. The total number of conditional production curves rises as  $\frac{1}{2}n(n-1)$ . The maximum number of conditional production curves which can support the budget hyperplane rises, however, as either  $\frac{1}{4}n^2$  (when  $n$  is even and when the relative wage rate makes one half of the input rays more capital intensive than the endowment ray) or as  $\frac{1}{4}(n+1)(n-1)$  [when  $n$  is odd and  $\frac{1}{2}(n+1)$  rays are either more or less capital intensive than the endowment ray]. For a large  $n$ , therefore, the ratio of supporting to total conditional production curves is approximately one half while, of course, the production surface and the budget hyperplane have the same dimensionality.

Notice that because the tangency of the

budget planes extends all the way to the relevant two-dimensional axial planes it is incorrect to think of the production surface as consisting of flat facets separated by what would be, in the two-dimensional case, vertex points. The relationship between the line  $q_{13}q_{12}$  and the surface  $Q_1Q_2Q_3$  in Figure 3 provides a better, if still incomplete, image the shape of the tangency region in high dimensions where however, as we have seen, the tangency region and the production surface have the same dimensionality.

## II. Specialization

Consider next the complicated question of specialization. As we have seen above, every factor budget line  $Y = wL + rK$ , and therefore every relative wage rate, can be associated with a budget plane  $Y = \sum_{i=1}^n p_i x_i$  and therefore with a set  $p = (p_1, \dots, p_n)$  of  $(n-1)$  relative prices:

$$(2) \quad \begin{aligned} p_1 &= 1, \\ p_i &= \pi_i(\omega) \quad (i = 2, \dots, n) \end{aligned}$$

consistent with the relative wage rate  $\omega$ . Here the first product, perfectly arbitrarily, has been called the numeraire product.

On the other hand, it takes only one relative price to determine  $\omega$  (as we know from the factoral-price equalization theorem) so long as one of the corresponding capital-labor ratios exceeds the other at all relative wages. Equations (2) thus impose on product prices a consistency relationship which may limit the number,  $n'$ , of products which the domestic economy, facing foreign competition at the internationally given prices, can produce without loss.

The number  $n'$  of mutually consistent prices is a convenient measure of the degree of specialization and can be determined as follows. The international prices and equation (1) identify, for each product, the isoquant of which the output

is worth  $\$Y$  (where  $Y$  is arbitrary). This can be called the  $\$Y$  product isoquant of product  $i$ . The resulting  $n$   $\$Y$  product isoquants form in factor space a convex hull which can be regarded as the isoquant for producing  $\$Y$  of foreign exchange at the international prices. By adjusting  $Y$ , one can find the convex hull which contains the endowment point  $E$  and thus find the highest attainable value,  $\$Y^*$ , of national income.

The slope of the convex hull at  $E$  determines the competitive value of  $\omega$  and thus the slope of the  $\$Y^*$  factor budget line. Any other line segment through  $E$  would intersect at least one  $\$Y^*$  product isoquant and thus permit, contrary to the assumption of competition, positive profits to be earned indefinitely. The production of any product of which the  $\$Y^*$  product isoquant fails to share at least one point with the  $\$Y^*$  factor budget line incurs losses, meanwhile. Clearly therefore  $n'$  can lie anywhere between one and  $n$ , depending on the configuration of prices facing the economy. Fortunately and obviously everything stated in the preceding section for the case of  $n' = n$  products holds also for that of  $0 \leq n' \leq n$ .

## III. Indeterminacy and the Theory of Protection and Trade

It can be seen from the preceding two sections that the exact output of any product is indeterminate under normally observed conditions and therefore extremely sensitive to relative prices. It would be singular if the rest of the world presented the economy with a full set of  $n$  consistent prices, or even of very many,  $n'$ , such prices. The home (country  $A$ ) authorities can always make  $n'$  as large as desired by setting tariffs,  $t_i^A$ , of the required levels. Because, by equation (2), prices in each country depend on its own relative wage rate, the equilibrium tariffs necessary to achieve a given value of  $n'$

will be functions of *both* national relative wage rates,  $\omega^A$  (country *A*) and  $\omega^B$ . Equilibrium tariff rates can be defined as those which just equate foreign and domestic costs, in both countries. If we let  $c_i$  stand for the ratio of *A*'s cost of producing the *i*th product to *B*'s cost of producing the same product, then we can define the equilibrium, or *minimum-prohibitive*, duties,  $m_i^A$  and  $m_i^B$  affecting product *i* by the relationships:

$$\frac{(m_i^A + 1)}{(m_i^B + 1)} = c_i$$

$$m_i^A m_i^B = 0$$

where  $m_i^A$  is the proportion of the price of product *i* in country *B* which is collected (by either or both authorities) if  $c_i$  exceeds unity (and *A* imports the product) and  $m_i^B$  is the proportion of the price of product *i* in country *A* when unity exceeds  $c_i$ .

Note that both  $m_i^A$  and  $m_i^B$  cannot differ from zero. For this reason we can drop the country superscript in general discussion and speak simply of  $m_i$ , the minimum-prohibitive duty as applied in the relevant country. Then because  $c_i$  is a function only of the two relative wage rates:

$$(3) \quad c_i = \phi_i(\omega^A, \omega^B) \quad (i = 1, \dots, n)$$

so is  $m_i$ .

Under the above formulation, a tariff rate  $t_i^A < m_i^A$  set, in this case by country *A*, establishes a domestic price of the *i*th product inconsistent with the relative wage rate established by the  $n'$  consistent duties. The domestic industry then declines without in the least affecting either relative wage rate or the mutual consistency of the  $n'$  prices of products enjoying their minimum prohibitive duties.

If any consistent duty is *raised*, by contrast, it immediately determines a new subset of only one or two mutually consistent prices. If the overprotected (rela-

tive to the previous relative wage rates) products are sufficiently important, the remaining domestic industries must decline or else receive higher duties. The new duties set the same process in operation, though with respect to another set of industries. The basic asymmetry between the situations  $t_i^A < m_i^A$  (which causes the *i*th industry to shrink without lowering  $\omega^A$ ) and  $t_i^A > m_i^A$  (which immediately raises  $\omega^A$  and renders inconsistent the duties on other products) evokes adjustments which, if they overshoot the mark, keep raising  $\omega^A$  (and depressing  $\omega^B$ ) until finally these rates reach their autarkic levels. At that stage, trade in "Heckscher-Ohlin" products (of which equation (1) fairly represents the production functions and factor inputs) has been eliminated.

The indeterminacy of outputs can start this process even if consistent duties are set for each intermediate level of  $\omega^A$  and  $\omega^B$ . This "driftiness" can be illustrated in the geometrical, three-commodity illustration of Figure 3, although its force is even greater in higher dimensions. Suppose that the budget hyperplane shifts slopes in such a way that it always remains in contact with defining lines (such as  $q_{13}q_{12}$ ) of the production surface. As we have seen, it then remains in contact with the principle conditional production curve,  $Q_1Q_n$ . Let the shift in slope be such that the output point  $q_{13}$  moves downward, in the direction of the small arrow adjacent to it in Figure 3. The entire defining line  $q_{13}q_{12}$  moves downward then, as indicated by the two arrows at its endpoints. The corresponding "rotation" of the consistent budget plane implies a rise in the prices of the third and second products, because the intersections of the budget plane with the corresponding commodity axes move inward, as shown by the adjacent arrows. The rotation also implies, of course, a drop in the price of the first commodity. At a given moment, however, the rotation

of the consistent budget plane (corresponding to the rotation of the factoral budget line  $X_3X_2EX_1$  in Figure 2) will imply a fall in the price of the second commodity, as its input ray  $OX_2$  in Figure 2 crosses the endowment ray, and it becomes, like the first commodity, more labor intensive than the economy (as defined by the endowment ray  $OE$ ).

At any moment during this rotation, however, the actual output vector can shift from the principle conditional production curve to one of the others, or back again. Thus, as the price of the third product progressively rises until the defining line  $q_{13}q_{12}$  reaches the point  $Q_2$ , the output of the third commodity could indifferently trend toward zero as  $d$  in the expression

$$dq_{13} + (1 - d)q_{12} = y$$

trended toward zero. Once the endpoint  $q_{12}$  reaches  $Q_2$ , of course, and switches to a point  $q_{23}$ , the output of the third product must then begin to rise as  $q_{23}$  moves along the conditional production curve (2, 3).

The above consideration bears on the use of effective rates of protection, the proponents of which claim to be able to infer the direction in which the tariff shifts resources (see W. M. Corden, p. 266) at least if all production functions are known (see Roy Ruffin, p. 268). Unfortunately the usual assumptions (first-degree homogeneity and many goods) of effective protection theory imply indeterminacy whenever there are few factors; either the assumption of a factor specific to each product or the assumption of decreasing returns to scale would be required to rule it out.

The indeterminacy of the structure of production, and therefore of trade, washes out, as Melvin indicates (p. 1284), when one looks at the underlying balance of *factoral* trade (see also Travis, pp. 140-43), however. The milling around of domestic

resources in response to slight price inconsistencies is harmless in this system, moreover, because there is no producers' surplus, other than the equilibrium returns to capital and labor, to be gained or lost (see E. J. Mishan (1968 and 1971)).

Clearly, however, factoral trade is stopped if the process of setting tariffs piecemeal, in response to drift or to accidental inconsistency in the structure of rates, causes all tariffs finally to exceed their corresponding minimum prohibitive levels. This is serious because capital and labor each account for about half of national income (when capital is defined to include human capital) and are very unevenly distributed internationally. The United States, for instance, probably has about one third of the world's capital supply, yet only one twentieth of its labor supply. If the tariff stops trade in products (generally manufactures and services) of which equation (1) adequately describes the production function (as opposed to oil, copper, lead, zinc, and so forth which account for a high percentage of present U.S. trade), it may be quite costly.

There is abundant empirical evidence now to support the above viewpoint. Wassily Leontief (1953, 1956) discovered that in 1949 and 1953, U.S. imports of manufactures were no more labor intensive than its exports of manufactures, a finding which Robert Baldwin has confirmed with more recent data. Gary Hufbauer, pp. 14, 44-51, 96 and 97, has found that other countries with extreme relative endowments of labor and capital fail to exchange labor-intensive for capital-intensive products. Yet rankings of manufactures by labor-intensiveness are internationally stable, as Hufbauer, p. 14, and Hal Lary have shown. Distinct input rays, such as those which Figures 1, 2, 4, and 5 depict, can be posited, therefore. The remaining empirical support which the above model

of protection requires is to see if duties are high enough to eliminate what might be called Heckscher-Ohlin trade.<sup>1</sup>

#### IV. U.S. Duties Compared with their Minimum-Prohibitive Levels

In order to compute  $m_i$ , the tariff rate which would just shut off competitive imports of the  $i$ th product, equation (3) must be solved. This equation derives from equation (1) and the marginal-productivity conditions. In deriving (3), it is convenient to define units so that  $A_i$ , the neutral efficiency multiplier, is equal to unity in country  $B$  and thus indicates country  $A$ 's efficiency, relative to that of country  $B$ , in producing the  $i$ th product. Then, if product one is taken as the numeraire (and has the same price in both countries),  $c_1 = 1$  and<sup>2</sup>

$$(4) \quad c_i = \frac{A_1 \psi_1(\rho_1^A) \psi_i(\rho_i^B) (\omega^A + \rho_1^A) (\omega^B + \rho_1^B)}{A_i \psi_1(\rho_1^B) \psi_i(\rho_i^A) (\omega^A + \rho_1^A) (\omega^B + \rho_i^B)} \quad (i = 2, \dots, n)$$

<sup>1</sup> The United States exchanges products which are somewhat more capital-intensive than the economy for products which are somewhat more capital-intensive still. Baldwin examines the many alternative explanations that have been supplied to explain this paradoxical structure of trade and very carefully and critically lays bare the logical and empirical issues involved. Using a different method from that used here, he too concludes, p. 143, that protection affects the pattern and factor-content of trade.

<sup>2</sup> The derivation (see Kemp, pp. 7, 8) is:

$$p_i = r L_i (\omega + \rho_i) / x_i$$

(under competition price equals total factor cost) and therefore

$$p_i / p_1 = \frac{L_i x_1 (\omega + \rho_i)}{x_1 L_1 (\omega + \rho_1)}$$

From (1) it follows that

$$L_i / x_i = 1 / A_i \psi_i(\rho_i)$$

and that, in country  $A$

$$\frac{p_i^A}{p_1^A} = \frac{p_i^A}{p_1^B} = \frac{A_1 \psi_1(\rho_1^A) (\omega^A + \rho_i^A)}{A_i \psi_i(\rho_i^A) (\omega^A + \rho_1^A)}$$

Because the capital-labor ratios

$$(5) \quad \rho_i = \rho_i(\omega) \quad (i = 1, \dots, n)$$

are functions only of the relative wage rate, they can be eliminated in (4), although doing so requires prior knowledge of the form  $\psi_i(\rho_i)$  in (1). This form governs, however, the  $i$ th product's elasticity,  $\sigma_i$ , of factoral substitution, which must exceed zero. When it is zero, the  $\rho_i$  are constants independent of the relative wage rate and are therefore internationally identical. The forms  $\psi_i(\rho_i)$  are also internationally identical, so that one usable specification of (4) is

$$(6) \quad c_i = \frac{A_1 (\omega^A + \rho_1) (\omega^B + \rho_1)}{A_i (\omega^A + \rho_1) (\omega^B + \rho_i)} \quad (i = 2, \dots, n)$$

Although in empirical work the substitution elasticity is often assumed, for convenience, to be zero, this is obviously a limiting case. By assuming that equation (1) exhibits for all values of  $\rho_i$  a constant elasticity of substitution (*CES*), a specialization of equation (4) good for all assumed values (except unity) of that elasticity can be derived. The internationally identical form in (1) can then be written as

$$(7) \quad \psi_i(\rho_i) = [(1 - \lambda_i) \rho_i^{(\sigma_i - 1) / \sigma_i} + \lambda_i]^{(\sigma_i / (\sigma_i - 1))} \quad (i = 1, \dots, n)$$

where the distribution parameters  $\lambda_i$  are internationally identical constants (see Kenneth Arrow, Hollis Chenery, Bagicha Minhas, and Robert Solow). The marginal productivity condition,

$$\omega = \frac{\psi_i(\rho_i) - \rho_i \psi_i'(\rho_i)}{\psi_i'(\rho_i)},$$

Because a similar expression obtains for country  $B$ , equation (4) gives an expression for

$$c_i = \frac{p_i^A p_1}{p_1^A p_i^B}$$

$$(8) \quad c_i = \frac{A_1(\rho_1^A + \omega^A)^{1/(\sigma_1-1)} [\rho_1^A \omega^B(\sigma_1-1) + \omega^A \sigma_1]^{1/(\sigma_1-1)}}{A_i(\rho_i^A + \omega^A)^{1/(\sigma_i-1)} [\rho_i^A \omega^B(\sigma_i-1) + \omega^A \sigma_i]^{1/(\sigma_i-1)}}$$

the assumption of competition, and equation (7) further imply that in equation (5)

$$\rho_i(\omega) = \omega^{\sigma_i} \left( \frac{1 - \lambda_i}{\lambda_i} \right)^{\sigma_i}$$

Further, using equation (4), and assuming  $c_1=1$ , we obtain equation (8).

Empirical studies have estimated the substitution elasticities of a number of sectors. The overall impression (see Marc Nerlove and J. C. R. Rowley) they leave is that most estimates lie below, but few are significantly different from, unity. When  $c_1=\sigma_1=1$ , equation (4) specializes to

$$c_i = \frac{A_1}{A_i} \left( \frac{\omega^A}{\omega^B} \right)^{(\lambda_i - \lambda_1)}$$

where the distribution parameter  $\lambda_i$  ( $i=1, 2$ ) is labor's share of the price of the  $i$ th product, and, in terms of the more easily observed domestic capital-labor ratios, to

$$(9) \quad c_i = \frac{A_1}{A_i} \left[ \frac{\omega^A}{\omega^B} \right]^{\{(\omega^A(\rho_1^A - \rho_i^A)) / (\omega^A + \rho_1^A) - (\omega^A + \rho_i^A)\}}$$

Table 1 gives estimates of the  $\rho_i^A$  for a small sample of the nearly 200 Input-Output product categories. It gives, in columns 3 and 4, the estimates of  $c_i$  made with equations (6) and (9), respectively, and thus reflecting the assumptions  $\sigma_1=\sigma_i=0$ , on the one hand, and  $\sigma_1=\sigma_i=1$ , on the other. In both columns 3 and 4 the *U.S.* national product, or *NNP*, is taken to be the numeraire (product one). The resulting values of  $c_i$  are thus easily compared, and the assumption that  $\sigma_1=\sigma_i=1$  is seen to imply the higher disparity  $c_i/c_1$ , for any numeraire common to both computations.

The costs  $c_i$  can be compared with the

price disparities,  $(1+t_i^A)$ , imposed by the *U.S.* tariffs,  $t_i^A$ , on all *U.S.* imports. These are reported in column 5. The sample includes *all* product categories for which  $t_i^A$  is less than twice as high as  $(c_i-1)=m_i^A$ , the *U.S.* minimum-prohibitive duty on the  $i$ th product.<sup>3</sup> It can be seen that only 12 categories fail to receive minimum-prohibitive duties and that most of these, like sector 187 (business services), may be objectively difficult to trade and their trade may be difficult to tax. Because the assumption that all  $\sigma_i=\sigma_1=1$  implies the greater minimum-prohibitive duties, we can conservatively base the discussion of the restrictiveness of the tariff upon that assumption in discussing Table 1.

While the relative values of the  $c_i$  are independent of the choice of numeraire, their absolute values are not. Relative and absolute values of the  $c_i$  both depend, moreover, on  $A_1$  and  $A_i$ , the international relative efficiency coefficients. In order to isolate the Heckscher-Ohlin trade basis, Table 1 assumes that all  $A_i=A_1=A$ , where  $A$  is therefore the common inter-industrial factor by which *U.S.* technical efficiency differs from that of the rest of the world.

Finally, Table 1 requires data on  $\omega^A$  and

<sup>3</sup> The complete table will be supplied by the author upon request. The most salient feature of the full table which the sample fails to convey is that even most products for which  $c_i < 1$  enjoy substantial *U.S.* duties, even though they are presumably *U.S.* export products. Several explanations, in addition to the general over-protectiveness of the tariff, can be adduced. Some capital-intensive products (e.g., textiles, which the *U.S.* imports despite a high tariff) use even more capital-intensive intermediate inputs (e.g., cotton, which the United States exports freely) to imports of which foreign countries may fail to apply minimum-prohibitive duties. In such cases, formula (4) must be altered to calculate the revised *U.S.* minimum-prohibitive duty.

TABLE 1—SELECTED U.S. CAPITALIZATION INDEXES, COMPARATIVE COSTS, AND  
TARIFF-IMPOSED INTERNATIONAL PRICE DISPARITIES<sup>a</sup>

Product's Input-Output Number and Name: (1)	Total Capitalization Index, $\rho^A$ : (2)	Comparative Costs, $c_i$ , when:		Tariff-Imposed International Price Disparities, $(1+t_i^A)$ : (5)
		$\sigma_i=0$ : (3)	$\sigma_i=1$ : (4)	
167 Electric light & power	311.64	0.70378	0.55287	1.00
5 Cotton	259.04	0.73387	0.60014	1.00
30 Spinning, weaving, dyeing	115.54	0.94424	0.92021	1.240
156 Watches, clocks	100.01	0.99999	0.99996	1.50
135 Electrical appliances	99.85	1.00061	1.00001	1.175
136 Advertising, incl. radio, television	94.19	1.02507	1.03450	1.00
16 Coal mining	92.87	1.03160	1.04323	1.00
134 Laundries, dry cleaning	91.37	1.03832	1.05220	1.00
135 Other personal services	91.29	1.03870	1.05386	1.065
144 X-ray apparatus	90.97	1.04022	1.05595	1.075
67 Leather tanning & finishing	90.72	1.04144	1.05762	1.10
138 Electric lamps	88.54	1.05233	1.07253	1.12
178 Local & highway transportation	88.07	1.05475	1.07582	1.00
148 Aircrafts & parts	86.62	1.06234	1.08615	1.125
131 Motors, generators	86.34	1.06380	1.08812	1.125
141 Communication equipment	85.95	1.06590	1.09095	1.175
110 Steam engines, turbines	85.72	1.06717	1.09265	1.15
188 Automobile repair services, garages	84.77	1.07231	1.09961	1.095
160 Office supplies	84.69	1.07275	1.10019	1.15
38 Plywood	84.22	1.07536	1.10371	1.179
137 Engine electrical equipment	82.87	1.08300	1.11395	1.10
41 Wood furniture	81.32	1.09208	1.12605	1.143
39 Fabricated wood products, excl. furn.	81.16	1.09301	1.12729	1.15
117 Cutting tools, jigs, fixtures	81.03	1.09381	1.12835	1.25
139 Radio & related products	80.22	1.09869	1.13485	1.12
163 Motion-picture production	79.45	1.10341	1.14105	1.05
40 Wood containers, cooperage	78.82	1.10732	1.14622	1.05
158 Musical instruments & parts	78.26	1.11089	1.15092	1.20
189 Other repair services	77.00	1.11899	1.16153	1.00
68 Other leather products	72.29	1.15158	1.20364	1.20
181 Banking, finance, insurance	70.17	1.16749	1.22383	1.00
69 Nonrubber footwear	70.09	1.16809	1.22461	1.166
187 Business services	64.21	1.21716	1.28559	1.00

<sup>a</sup> Sources of data and methods of computation are found in the Appendix to this article.

$\omega^B$ . These and the value of  $A$  (which does not affect relative costs, of course) were estimated and the choice of numeraire were made on the basis of the assumptions that U.S. income is one quarter of world income, U.S. labor employment is one twentieth of world labor employment, the average rate of return to capital is .100 in the United States and .15833333 elsewhere, and that the aggregative national-income production functions can be written as

$$(10) \quad NNP = AL\psi(\rho) = A\sqrt{LK}$$

In equation (10)  $K$  includes all forms of human capital as well as tangible capital and units are so chosen that  $A$  equals unity in country  $B$  and  $\rho^A=100$ . These assumptions and units imply that

$$\omega^A = \rho^A = 100,$$

$$\omega^B = 9.97229916,$$

and that  $A$  equals two.

Because in most countries the product of the ruling unskilled wage rate and the total labor input equals just one half of *NNP*, equation (10) is probably a reasonable specification of the production function of the national product. It implies, of course, that the elasticity of factoral substitution in producing that product is unity, and so the national product itself can reasonably be taken as the numeraire when the substitution elasticities of all separate products are taken to be unity as well. Thus Table 1 takes any product for which the capital-labor ratio is equal to 100 as the numeraire. The choice of units renders as index numbers therefore all the other capital-labor ratios reported there.

Fortunately when all  $\sigma_i = \sigma_1 = 1$  the only parameter we need in equation (9) to compute the  $c_i$  for a given choice of numeraire is the relative wage rate,  $\omega^A/\omega^B$ . The various assumptions above imply that this parameter equals 10.027778 which, compared with direct observations of wages and profit rates in the United States and other countries, seems conservative enough.

It is necessary, nevertheless, to test the results of Table 1 for their sensitivity to even greater international disparities in relative factor costs and to free, as far as possible, their interpretation from the choice of numeraire and of common substitution elasticity.

#### V. Sensitivity of Comparative Costs to Substitution Elasticities and Relative Wages

In addition to showing all tariffs that are less than twice their minimum-prohibitive levels, Table 1 also reports the sector (category 187) which needs the highest duty of all the minimum-prohibitive rates in order to shut off *U.S.* imports. This rate is 28.559 percent. By the same token, the highest minimum-prohibitive rate which country *B* requires to stop all competitive imports is 80.741 percent (for

category 167), corresponding to the maximum capitalization index, 311.64, which Table 1 (which ranks by capital-intensiveness) reports. Presumably, therefore, a *U.S.* tariff rate of about 30 percent on all products, combined with a foreign tariff of about 80 percent, would stop all *U.S.* competitive trade, and all duties but those two would be redundant.

The minimum-prohibitive duties of a single country cannot be assessed accurately without knowing which numeraire to select. There exists, however, a relationship between the highest minimum-prohibitive duties of any two countries which is invariant to the choice of numeraire: namely their sum plus their product. If we let  $M^A$  and  $M^B$  stand, respectively, for *A*'s and *B*'s maximum minimum-prohibitive duties then, for any given numeraire,

$$\text{Max}_{i=1}^n c_i = (1 + M^A),$$

$$\text{Min}_{i=1}^n c_i = 1/(1 + M^B),$$

and therefore, for any common numeraire,

$$\text{Max}_{i=1}^n c_i / \text{Min}_{i=1}^n c_i = 1 + (M^A + M^B + M^A M^B)$$

In particular, if we number products from 1 to  $n$  by capital intensiveness, then  $c_n/c_1$  is the maximum comparative cost that can be obtained by any choice of numeraire and

$$(11) \quad (M^A + M^B + M^A M^B) = \left( \frac{c_n}{c_1} - 1 \right)$$

This relationship provides a convenient way to assess the joint restrictiveness (of their bilateral trade) of any two countries' tariff schedules.

Table 2 illustrates the method by using equations (8) and (9) to compute the maximum comparative-cost ratio,  $c_n/c_1$ . The numeraire product is category 167 of Table 1 and product  $n$  is category 187 of that table. These products have, respectively,

the highest and lowest capitalization ratios of all 190 Input-Output product categories.

Table 2 reports  $c_n/c_1$  for the hypothesis that  $\omega^B = 9.9722916$  and for the hypothesis that  $\omega^B$  is only half that amount, or 4.98614958. This latter hypothesis implies a relative-relative wage rate of more than twenty, and probably represents fairly therefore even the vast disparity between, say, U.S. and Indian relative wage rates. The rows of Table 2 correspond to several alternative hypotheses regarding the value of the factoral-substitution elasticity,  $\sigma$ , assumed to be invariant with respect to the relative wage rate and common to both product categories. The selection of values of  $\sigma$  includes  $\sigma = 1.2255$ , for which  $c_n/c_1$  reaches its maximum when  $\omega^B = 9.9722916$ , and  $\sigma = 1.1734$ , for which  $c_n/c_1$ , equal to 3.06184, reaches its maximum for all values of  $\omega^B$  which exceed 4.98614958.

Even when  $c_n/c_1$  reaches this maximum, equation (11) indicates that a maximum U.S. minimum prohibitive duty of, say, .40 (or 40 percent) combined with a maximum foreign minimum-prohibitive duty of 1.9 (or 190 percent) would stop all bilateral trade in question.<sup>4</sup>

We conclude, therefore, that the tariff is a sufficient explanation of the failure of the United States to exchange capital-intensive products for labor-intensive products even with such abundant-labor countries as India.<sup>5</sup> Leontief's 1953 and 1956 studies

TABLE 2—MAXIMUM COMPARATIVE COSTS,  $c_n/c_1$ , FOR SELECTED ELASTICITIES OF SUBSTITUTION AND VALUES OF  $\omega^B$  WHEN  $\omega^A = 100$ ,  $p_1^A = 311.64$ , AND  $p_n^A = 64.21$

Elasticity of Substitution, $\sigma = \sigma_1 = \sigma_n$ :	Country B's Relative Wage Rate, $\omega^B$ :	
	$\omega^B = 9.9722916$	$\omega^B = 4.98614958$
0.00	1.72947	1.82535
0.25	1.87647	2.05947
0.50	2.04303	2.37000
0.75	2.20577	2.72177
1.00	2.32534	2.99693
1.1734	2.36187	3.06184
1.2255	2.36396	3.05597
1.25	2.36350	3.04909
1.50	2.30920	2.85644
1.75	2.18643	2.54384
2.00	2.03490	2.23912
3.00	1.56427	1.57844
4.00	1.35733	1.35830

(which were based on essentially the same production data as is Table 1) and the recent studies of Hufbauer and Baldwin confirm this conclusion by showing that little Heckscher-Ohlin trade takes place.

By the same token, the tariff explains the coexistence of incomplete specialization with vast international disparities in relative factor prices which, under Heckscher-Ohlin assumptions including freedom of trade, would represent a contradiction. Peter Kenen has asserted, p. 439, for example that tariffs cannot explain observed factor-price disparities, and he presents a new theory to explain those disparities under free trade. It appears now, even within the familiar theoretical framework, that the seemingly moderate duties are in fact prohibitive of trade even

<sup>4</sup> Jagdish Bhagwati and Padma Desai, in their comprehensive and authoritative study of Indian industrialization and trade policies since 1951, document Indian rates of protection of several hundred percent, pp. 281-367, combined with a highly overvalued currency. The overvaluation of the currency acts, of course, as an across-the-board export duty, which should be added to the U.S. duties in obtaining the initial estimates of the  $t_i^A$ . A large number of poor countries in addition to India maintain (with the aid of strict import and exchange controls) heavily overvalued currencies.

<sup>5</sup> The assumption that all  $A_i$  equal  $A$  underlies this conclusion, however, and is certainly counterfactual. Gary Bickel, using estimates of  $A_i$  (where Japan is country B) provided by Arrow, Chenery, Minhas, and

Solow, found that at least 25 percent of the total variation in U.S.-Japanese prices was attributable to the efficiency factor (see Baldwin, p. 128). This does not mean, of course, that the variance among the  $A_i$ 's defined for the entire rest-of-the-world was this great, while it does mean that other factors, notably the variance of the capitalization indexes, dominate the contribution made by the variation of the efficiency factors.

If transportation and other objective trading costs are important, the tariff may not be the sole explanation of the failure of the United States to exchange capital-intensive for labor-intensive products. In some cases the transportation costs alone obviously exceed the international cost disparities.

as between countries with widely different relative capital and labor endowments.

On second thought this is not astonishing, but is simply a consequence of what Ronald Jones calls the magnification effect, namely that changes in relative prices give rise to magnified changes in relative factor rewards. This effect is closely related, of course, to the theorem of Wolfgang Stolper and Samuelson that a tariff can raise a factor's real earnings even with respect to the protected product.

The reader may feel that the comparison of minimum-prohibitive duties with actual duties proves too much because, after all, some manufactures are traded. The many recent studies of the commodity-composition of trade seem to reveal, however, that much *U.S.* trade in manufactures consists of products requiring special skills, products recently introduced in response to peculiar *U.S.* incomes and relative prices, and products in which, for one reason or another, the United States undertakes an abnormal amount of research and development (see Hufbauer, Donald Keesing, Vernon, and William Gruber, Dileep Mehta, and Vernon). It is not difficult to see how such products might filter through the tariff, which is revised only from time to time. Its authors can set redundant rates for every product they have heard of, but not for those to come. In addition, some of the more dramatic increases in imports of manufactures (notably autos, shoes, etc.) have occurred only since 1962 and seem, in accordance with the argument of this paper, to reflect the reduction since that time of their tariffs below their minimum-prohibitive levels.

## VI. Summary

This paper suggests a practical method for working with multiproduct situations of pure barter exchange. The method is convenient only if two primary factors are included, but this feature is not terribly

restrictive in practice thanks to the overwhelming importance of capital and labor compared to all other primary factors.

We have found, however, that the  $n$ -product, two-factor case gives rise to general-equilibrium output responses that are somewhat surprising when one is used to the determinate two-product, two-factor model or even to the many-product, one-factor model of Frank Graham (which has no relative wage rate and can give rise to no embodied exchange of factoral services). In particular, because a price increase raises the relative wage rate while a decrease fails to lower it, tariff authorities operating on individual industry requests for protection are likely eventually to establish a completely prohibitive tariff.

Because Leontief and others have found that *U.S.* trade in manufactures exchanges no embodied capital services for labor services and because such trade is less than 2 percent of *U.S.* income, I have argued that tariffs might be virtually prohibitive. A comparison of the minimum-prohibitive duties necessary to overcome the international cost disparities based on a rough estimate of present *U.S.* and foreign relative wage rates strongly confirmed this conjecture.

## APPENDIX

### Sources and Methods Used in Preparation of Table 1

In column (2), the capitalization ratios  $\rho_i^A$  were computed on the basis of data supplied by Leontief (1956, pp. 403-07, cols. 3 and 4). To the ratio of direct-plus-indirect capital per man was added, regardless of sector, an allowance of \$6,000.00 to express the quantity of human capital employed together with nonhuman capital. This figure was chosen as follows. The average amount of nonhuman capital per man in the productive sectors (which excludes trading activities) in 1947, on the basis of Leontief's data, was \$6,739.07961. Multiplying the economy's

total labor input in any year by the current *unskilled* wage rate (regardless of the level of qualification of individual workers) yields a product equal to about one half of current *NNP*. Yet capital in Leontief's sense (which excludes human capital) earns about one quarter of *NNP*. It follows that human capital, in the very broadest sense which includes the value (not necessarily the cost) of on-the-job training and all other remunerative experience (notably education), earns about one quarter of *NNP* as well, and must therefore be worth about the same amount as the stock of nonhuman capital (see Theodore Schultz). No attempt was made to correct the human-capital allowance for interproduct differences in skill requirements, some of which wash out anyhow in the construction of the direct-plus-indirect input requirements. Unless human-capital requirements are positively correlated with nonhuman capital requirements, neglecting interproduct skill differences will not bias downward the estimates of minimum-prohibitive duties. The corrected ratios were divided by \$127.3907961 to convert them to index numbers based on  $\rho^A = 100$ , where  $\rho^A$  is the amount per worker of human plus nonhuman capital in all U.S. productive sectors.

In column 5, the rates  $t_i^A$  are equal to the reported nominal duty rates, divided by 100. Those rates were taken from *U.S. Customs Duties Annotated for Statistical and Reporting Purposes*, July 1962 edition. The main difficulty was to establish concordance between the thousands of duties there reported and the 190 or so product-classes of column (1). Care was taken to report in column (5) the lowest duty which an actual importer would have to pay for a representative member of each 1-0 product class.

# REFERENCES

- K. J. Arrow, H. B. Chenery, B. S. Minhas, and R. M. Solow, "Capital-Labor Substitution and Economic Efficiency," *Rev. Econ. Statist.*, Aug. 1961, 34, 225-50.
- R. E. Baldwin, "Determinants of the Commodity Structure of U.S. Trade," *Amer. Econ. Rev.*, Mar. 1971, 61, 126-46.
- J. N. Bhagwati and P. Desai, *India: Planning for Industrialization: Industrialization and Trade Policies since 1951*, London 1970.
- G. Bickel, *Factor Proportions and Relative Price Under CES Production Functions: An Empirical Study of Japanese-U.S. Comparative Advantage*, Stanford 1966.
- W. M. Corden, "The Structure of a Tariff System and the Effective Protective Rate," *J. Polit. Econ.*, June 1966, 74, 221-37.
- F. D. Graham, "The Theory of International Values Re-examined," *Quart. J. Econ.*, Nov. 1923, 28, 54-86.
- W. Gruber, D. Mehta, and R. Vernon, "The R&D Factor in International Trade and International Investment of U.S. Industries," *J. Polit. Econ.*, Feb. 1967, 75, 20-37.
- G. C. Hufbauer, "The Impact of Country and Commodity Characteristics on Trade in Manufactured Goods," Univ.-Nat. Bur. Econ. Res., Conference on Technology and Competition in International Trade, New York, Oct. 11-12, 1968.
- R. W. Jones, "The Structure of Simple General-Equilibrium Models," *J. Polit. Econ.*, Dec. 1965, 73, 557-72.
- D. B. Keesing, "The Impact of Research and Development in U.S. Trade," *J. Polit. Econ.*, Feb. 1967, 75, 38-48.
- M. C. Kemp, *The Pure Theory of International Investment and Trade*, Englewood Cliffs 1969.
- P. B. Kenen, "Nature, Capital, and Trade," *J. Polit. Econ.*, Oct. 1965, 73, 437-59.
- H. B. Lary, *Imports of Manufactures from Less Developed Countries*, Nat. Bur. Econ. Res., *Stud. in International Economic Relations*, Vol. 4, New York 1968.
- W. W. Leontief, "Domestic Production and Foreign Trade: The American Capital Position Re-examined," *Proc. of the Amer. Philosophical Soc.*, Sept. 1953, 97, 332-49.
- , "Factor Proportions and the Structure of American Trade: Further Theoretical and Empirical Analysis," *Rev. Econ. Statist.*, Nov. 1956, 38, 386-407.
- J. R. Melvin, "Production and Trade with Two Factors and Three Goods," *Amer. Econ. Rev.*, Dec. 1968, 58, 1248-68.
- E. J. Mishan, "What is Producer's Surplus?" *Amer. Econ. Rev.*, Dec. 1968, 58, 1269-82.

- , "Mishan on the Gains from Trade: Reply," *Amer. Econ. Rev.*, Mar. 1971, 61, 202-07.
- M. Nerlove, "Recent Empirical Studies of the C.E.S. and Related Production Functions," in M. Brown, ed., *The Theory and Empirical Analysis of Production*, Nat. Bur. Econ. Res. Stud. in Income and Wealth, Vol. 31, New York 1967, 55-122.
- J. C. R. Rowley, "Investment Functions: Which Production Function?" *Amer. Econ. Rev.*, Dec. 1970, 60, 1008-12.
- R. J. Ruffin, "Tariffs, Intermediate Goods, and Domestic Protection," *Amer. Econ. Rev.*, June 1969, 59, 261-69.
- P. A. Samuelson, "Prices of Factors and Goods in General Equilibrium," *Rev. Econ. Stud.*, 1953-1954, 21, 1-20.
- W. A. Stolper and P. A. Samuelson, "Protection and Real Wages," *Rev. Econ. Stud.*, Nov. 1941, 9, 65-83.
- K. M. Savosnick, "The Box Diagram and the Production Possibility Curve," *Ekonomisk Tidskrift*, Sept. 1958, 60, 183-97.
- T. W. Schultz, "Reflections on Investment in Man," *J. Polit. Econ.*, Oct. 1962, supp., 70, 1-8.
- W. P. Travis, *The Theory of Trade and Protection*, Cambridge, Mass. 1964.
- R. Vernon, "International Investment and International Trade in the Product Cycle," *Quart. J. Econ.*, May 1966, 80, 190-207.
- U.S. Tariff Commission, *U.S. Customs Duties Annotated for Statistical and Reporting Purposes*, July 1962.

# Experimental Evidence on Alternative Portfolio Decision Rules

By M. J. GORDON, G. E. PARADIS, AND C. H. RORKE\*

The Markowitz model for discovering efficient portfolios, those for which the expected value of the return is maximized for each level of risk, has generated a considerable body of literature on the relation between return and risk for securities and portfolios. Perhaps the most important theorem to come out of this research may be summarized as follows. Let the expected value and standard deviation of the rate of return on an efficient portfolio of shares be  $\bar{k}_m$  and  $\sigma_m$ . William Sharpe, John Lintner, and Jan Mossin have shown that under plausible assumptions, if another portfolio with risk  $\sigma_j$  is efficient, its return will be

$$(1) \quad \bar{k}_j = i + \frac{\sigma_j}{\sigma_m} (\bar{k}_m - i)$$

where  $i$  is the interest rate on a risk free asset. It follows that a portfolio with return  $\bar{k}_j$  and a risk  $\sigma_j$  can be obtained by putting the fraction  $\lambda = \sigma_j / \sigma_m$  of wealth in the share portfolio  $m$  and  $1 - \lambda$  in the risk free asset. The selection of an optimal portfolio may therefore be reduced to the allocation of wealth between an efficient risky asset and a risk free asset.

The portfolio an investor considers optimal is considered a function of his risk aversion, and John Pratt has proposed the following basis for classifying utility functions according to their degree of risk aversion. If the *amount* invested in the risky asset increases (decreases) with wealth, the investor has decreasing (in-

creasing) absolute risk aversion. If the *fraction* of wealth invested in the risky asset increases (decreases) with wealth, the investor has decreasing (increasing) relative risk aversion.

James Tobin has demonstrated that to obtain utility indifference curves exhibiting risk aversion and yielding unique solutions to the above choice between a risky and riskless asset, it is sufficient to define the utility of wealth function of the individual to be monotonically increasing and strictly concave. Nils Hakansson, Henry Latané and Donald Tuttle, M. Freimer and M. Gordon, Edwin Neave, Menahem Yaari and others have explored the implications of alternative utility functions for an investor's portfolio decision. However, there has been no empirical research on optimal portfolio policy.<sup>1</sup> This may be attributed to the difficulty of obtaining data on investor portfolios and the truly formidable task of analyzing such data to obtain the relevant parameters for an investor's preferences.

An alternative to real data is the data that may be obtained from experimental games designed to eliminate some of the problems in obtaining relevant statistics from the data of actual portfolios. This paper describes such a portfolio game. It is designed to provide information on how the allocation of wealth between a risk free and a risky asset varies with the level of wealth. Section I describes the game.

<sup>1</sup> The experimental research with small gambles the objects of choice has practically no relevance for portfolio policy, since people do not make decisions with respect to their entire wealth in the same way that they react to trivial gambles.

\* The authors are on the faculties of the University of Toronto, Laval University, and Queen's University, respectively.

Section II describes the participants, the data of the experiment and other related information. Sections III and IV use the data to test the validity of alternative utility functions as descriptions of behavior.

### I. The Portfolio Game

The portfolio game, Aipotu (Utopia spelled backwards), may be summarized as follows. In Aipotu an individual's only source of income is his wealth. Periodically an Aipotuan makes a consumption decision, the amount of which is subtracted from his wealth, and an investment decision. The latter involves selecting one and only one among the five investment alternatives (gambles) listed in Table 1 and deciding how much money to invest in the gamble.

TABLE 1—INVESTMENT ALTERNATIVES IN AIPOTU

Gamble Number	Payoffs <sup>a</sup>		Probability	
	Red	Black	Red	Black
1	\$ 1.30	\$.80	.5	.5
2	1.50	.70	.5	.5
3	1.90	.40	.5	.5
4	2.50	.00	.5	.5
5	100.00	.00	.005	.955

<sup>a</sup> Amount investor receives per dollar played.

To illustrate the game's operation assume that an individual with wealth  $W_t = \$115,000$  at the start of period  $t$  decides to consume  $C_t = \$10,000$  and invest or play  $G_t = \$70,000$  on gamble 3. In return for the  $\$70,000$  he will receive  $P_t$ , which will be  $\$133,000$  or  $\$28,000$ . Hence his wealth at the start of  $t+1$  will be

$$\begin{aligned}
 W_{t+1} &= W_t - C_t - G_t + P_t \\
 &= \$115,000 - \$10,000 - \$70,000 \\
 &\quad + \$133,000 = \$168,000 \\
 \text{or} \quad &= \$115,000 - \$10,000 - \$70,000 \\
 &\quad + \$28,000 = \$63,000
 \end{aligned}$$

both with equal probability. If the individual had decided instead to play  $\$300,000$  on gamble 1:

$$\begin{aligned}
 W_{t+1} &= \$115,000 - \$10,000 - \$300,000 \\
 &\quad + \$390,000 = \$195,000 \\
 \text{or} \quad &= \$115,000 - \$10,000 - \$300,000 \\
 &\quad + \$240,000 = \$45,000
 \end{aligned}$$

The  $\$300,000$  play on gamble 1 involved "borrowing"  $\$195,000$ . Borrowing is allowed only to invest, and the amount borrowed must be repaid immediately upon the determination of the investment's outcome. Furthermore, negative wealth is not allowed, so that borrowing is limited to the amount which would yield a zero wealth should the outcome be black (unfavorable). Borrowing and lending are at a zero interest rate.

In Aipotu a periodic consumption of  $\$5,000$  provides the bare necessities of life,  $\$10,000$  provides a modest but comfortable standard of living, and affluence begins at  $\$20,000$  per year. Aipotuan are required to spend at least  $\$5,000$  per period on consumption. If consumption and bad luck reduce an Aipotuan's wealth to zero, he is transferred to "Lower Depths" which is roughly comparable to living on welfare. A resident of Lower Depths is given  $\$3,000$  per period in consumer goods and  $\$300$  in cash which he may save or invest in any one of the five gambles in Table 1. Finally, each Aipotuan is expected to try to accumulate enough wealth so as to be able to provide his children with an adequate endowment of wealth for living in Aipotu.

Each participant in the game is given a description of the game that contains the above information, he is given some exercises to insure he understands the mechanics of the game, he is given an initial wealth, and he is urged to put himself in the position of an Aipotuan in making his consumption, choice of gamble and scale of play decisions.

It may be advisable to note here certain characteristics of the five gambles which become evident to the participants more or less rapidly over time. Gamble 5 is like a lottery in that there is a very small probability of obtaining a very large payoff, and the expected value of the payoff is \$.50 per dollar played. The other four gambles are all profitable with the fourth being most profitable per dollar played. The expected values of the payoffs per dollar played are \$1.05 on the first gamble, \$1.10 on the second, \$1.15 on the third, and \$1.25 on the fourth. However, the risk of a gamble increases with its rate of profit and anyone who prefers more wealth to less wealth will prefer the second to the other three gambles. To see this recall that on each gamble the two outcomes are equally probable, and that the upper limit on an investor's play is the amount that would leave him with zero wealth in the event black. The consequence is that for any feasible play on gamble 1, 3, or 4, there is a gamble on 2 that will leave the investor with the same wealth in the event black and a larger wealth in the event red. This is illustrated in Table 2 where an individual with wealth of \$100,000 is assumed to make a play on each gamble that would leave him with zero wealth in the event black.

Figure 1 compares the four profitable gambles on the basis of return and stan-

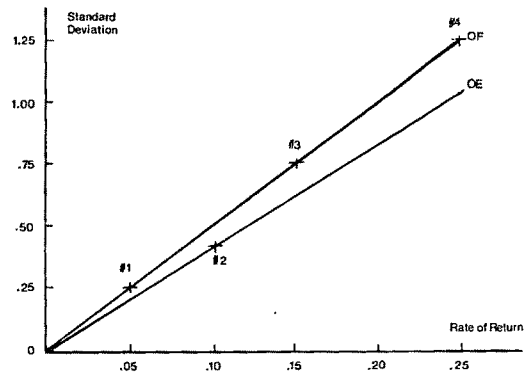


FIGURE 1

dard deviation. The return and standard deviation from investing all of one's wealth in gambles 1, 3, and 4 are the indicated points the line *OF*. Through the appropriate leverage or scale of play, any point on the line *OF* can be obtained with each of the gambles. The line *OE* represents the combination of return and risk that can be obtained by varying the scale on gamble 2.

The central purpose of the game would have been satisfied if the only gamble available to the participants was one among the four profitable gambles. The four were used to test the participants' ability to discover the conclusion presented in the last paragraph. Gamble 5 was included to test for the presence of risk lovers in making portfolio decisions.

A unique feature of Aipotu by comparison with most considerations of the portfolio problem is the stipulation that the individual consume a minimum amount as long as his wealth lasts. Without this stipulation a participant might reduce his periodic play and consumption to trivial levels and thereby avoid going broke indefinitely. This is unrealistic, since in real life an individual whose only source of income is his property or his property and labor jointly employed (a proprietorship) will not reduce his consumption below a subsistence level to avoid going broke. If

TABLE 2—OUTCOMES FOR MAXIMUM PLAY ON GAMBLES 1-4\*

Gamble Number	Play	Wealth	
		Red	Black
1	\$500,000	\$250,000	0
2	\$333,333	\$266,667	0
3	\$166,667	\$250,000	0
4	\$100,000	\$250,000	0

\* The investor's initial wealth is \$100,000, and he borrows the difference between his play and \$100,000 at a zero interest rate.

the individual is employable, he will prefer taking a job at a wage to continuing in business at the cost of a ridiculously low level of consumption. An individual who could not find employment would prefer going on welfare to living on his property income at a ridiculously low level of consumption. Notice the implicit but plausible assumption that being on welfare and having property income or that being an employee and having a proprietorship are mutually exclusive states. It is of course true that an individual can hold a job and own securities or other forms of property. Without modification Aipotu does not represent such a person, and such persons might behave differently than Aipotians.

## II. Behavior of the Participants

The participants in the experiment were thirty-four graduate students of business administration enrolled in a second course in finance, and the conduct of the experiment may be summarized as follows. Each participant was given an initial wealth that ranged from \$110,000 to \$190,000, and the class was divided into two equal groups, *A* and *Z*. Prior to a class meeting each student turned in his decisions for a period, and during the class a coin was tossed. If it turned up heads the outcome for group *A* participants was favorable and the outcome for group *Z* participants was unfavorable. The reverse took place if the coin turned up tails.<sup>2</sup> The toss determined his new wealth and the process was repeated at the next class meeting. At the end of eleven such trials the experiment was terminated without prior notice.

During the period of the experiment the course was devoted to the theory of security valuation and the cost of capital, including the influence of leverage on re-

TABLE 3—FREQUENCY DISTRIBUTION OF  
GAMBLES SELECTED

Gamble Number	1	2	3	4	5 <sup>a</sup>	0 <sup>b</sup>	Total
Frequency	10	253	20	78	11	2	374

<sup>a</sup> In every one of these cases the individual was in Lower Depths.

<sup>b</sup> Two individuals chose not to play on one trial.

turn and risk. Portfolio policy had not as yet been taken up, but a few students had taken another course which dealt extensively with the subject. Most of the students were finance majors and it was evident from their behavior that the game held their interest. They were highly motivated to succeed in Aipotu. To further motivate them the students were advised that at the termination of the experiment they would be required to write a short paper outlining what they considered the best strategy to follow in making consumption, gamble choice, and scale of play decisions.

By the time the experiment terminated practically all the students recognized that gamble 2 was the best. At the outset some selected gamble 1 because it had the lowest risk per dollar played and a fair number selected gamble 4 because it provided the highest return per dollar played. A few stayed with gamble 4 to the end. Gamble 5 was played only by individuals who had fallen into Lower Depths, and everyone reduced to that circumstance regularly played gamble 5. Therefore, none of the experiment's participants behaved like risk lovers as long as their wealth was large enough to provide some expectation of a livelihood under risk aversion behavior, and everyone deprived of that expectation became risk lovers. Table 3 presents the frequency with which each gamble was selected.

Table 4 presents the frequency of trials classified by wealth in intervals of \$25,000, the average wealth, the average consump-

<sup>2</sup> Another procedure was followed if the participant had selected gamble 5. A number from 1 to 200 was written on a piece of paper and he was called upon to guess what it was.

TABLE 4—RELATION BETWEEN CONSUMPTION,  $C$ , AND WEALTH,  $W$ 

$W$ Class	Frequency	Mean $W$	Mean $C$	Mean $C/W$
Less than \$25,000	14	\$ 18,471	\$ 5,100	.289
25,000- 50,000	23	38,539	5,270	.142
50,000- 75,000	22	61,186	5,527	.091
75,000-100,000	30	87,383	6,737	.077
100,000-125,000	58	112,374	7,900	.070
125,000-150,000	53	135,481	9,321	.069
150,000-175,000	48	162,440	9,531	.059
175,000-200,000	35	183,980	9,989	.054
200,000-225,000	12	212,117	10,792	.051
225,000-250,000	24	237,300	11,446	.048
250,000-275,000	11	261,045	11,400	.044
275,000-300,000	6	290,000	11,333	.039
300,000-325,000	10	311,810	12,160	.039
325,000-350,000	5	341,580	11,000	.032
350,000 and over	12	447,100	12,233	.028

TABLE 5—RELATION BETWEEN PLAY,  $G$ , AND WEALTH MINUS CONSUMPTION,  $W-C$ , ON GAMBLE 2

$W-C$ Class	Frequency	Mean $W-C$	Mean $G$	Mean $G/(W-C)$
Less than 25,000	16	\$ 15,750	\$ 38,594	2.51
25,000- 50,000	20	38,180	71,975	1.92
50,000- 75,000	20	62,360	81,850	1.31
75,000-100,000	31	88,503	93,332	1.07
100,000-125,000	39	112,077	96,428	.86
125,000-150,000	29	134,779	114,341	.85
150,000-175,000	33	162,852	126,827	.78
175,000-200,000	10	184,530	128,400	.70
200,000-225,000	9	212,067	121,778	.58
225,000-250,000	14	233,236	176,150	.76
250,000-275,000	6	263,083	149,167	.57
275,000-300,000	5	287,100	177,200	.61
300,000-325,000	6	309,333	159,767	.52
325,000-350,000	5	336,520	156,640	.47
350,000 and over	10	431,440	251,980	.58

tion and the average consumption-wealth ratio for each cell. Consumption rose with wealth, while the consumption-wealth ratio fell. The broad outline of the participants' consumption behavior, therefore, cannot be considered unreasonable.

The amount gambled by an individual depended among other things on the gamble selected, and some idea of the relation between scale of play and wealth is provided by the data on gamble 2 contained in Table 5. The table presents the frequency of plays on gamble 2 classified by wealth  $W$  minus consumption  $C$  in intervals of \$25,000, the average value of  $W-C$ , the average size of play  $G$ , and the average  $G/(W-C)$  ratio. The data for the other gambles do not convey a different impression. On average the amount gambled increased with  $W-C$ , but it cannot be ascertained from the data whether this is due to covariation between play and wealth for given participants or due to covariation between play and wealth among participants. There is also a clear tendency for the fraction of wealth played to decline as wealth increases, at least up to a wealth of \$200,000.

In writing the paper outlining what they

considered optimal policy with respect to consumption, gamble choice, and scale of play, the students were instructed not to rationalize the strategy they followed if they no longer considered it optimal. Nonetheless, it is of interest to review some of the policies they recommended for scale of play. A very common formulation of the problem may be summarized as follows. The periodic play must be small enough so that a few unfavorable outcomes do not result in bankruptcy. At the same time it must be large enough so that on an expected value basis the outcomes cover consumption and some increase in wealth. One strategy was to arrive at a periodic consumption and play such that  $n$  consecutive losses would be required for bankruptcy, where the probability of  $n$  was small enough to be tolerable.

The above strategy implied a constant consumption and play per period. Some students actually did that. Some followed this strategy until a significant change in  $W$  had taken place. A rise in wealth led to a rise in consumption and in the scale of play but both were kept to a level consistent with a larger value of  $n$ . A fall in wealth led to a fall in both variables.

However, it was recognized that the above strategy broke down as wealth became small. The risk associated with a large play in relation to wealth was necessary to avoid the high likelihood of Lower Depths under a conservative strategy. Another strategy was to set the scale of play so that the expected value of the outcome covered consumption and some increase in wealth. The consumption and the desired increase in wealth varied more or less with wealth from one period to the next. A few participants proposed varying the play inversely with wealth so as to repair the loss in wealth due to an unfavorable outcome and maintain a desired long-run growth in wealth.

### III. Alternative Utility Functions

In attempting to analyze the utility function implication exhibited by this game we assumed that multi-period behavior could be conceived as a sequence of one period decisions. That is, the individual, having made his consumption decision, would then proceed to invest and maximize his one-period investor utility with his gamble decision.

With this formulation in mind, five utility of wealth functions were formulated and tested. These are:

- 1) The quadratic function:

$$U(W) = W + \alpha/2W^2$$

- 2) The logarithmic function:

$$U(W) = \ln(W)$$

- 3) The power function:  $U(W) = W^\delta$

- 4) The adjusted logarithmic function:

$$U(W) = \ln(W+k)$$

- 5) The adjusted power function:

$$U(W) = (W+k)^\delta$$

where  $\alpha$  and  $\delta$  are risk preference parameters and  $k$  is a constant.<sup>3</sup>

TABLE 6—PRATT'S CLASSIFICATION OF UTILITY OF WEALTH FUNCTIONS

	Absolute risk aversion	Relative risk aversion
Quadratic	Increasing	Increasing
Logarithmic	Decreasing	Constant
Power	Decreasing	Constant
Adjusted Logarithmic	Decreasing	Increasing
Adjusted Power	Decreasing	Increasing

Table 6 presents an analysis of these functions according to the previously mentioned Pratt mechanism for describing the risk aversion characteristics of utility functions. As noted by Pratt and others there is an intuitive expectation that people will exhibit decreasing absolute risk aversion, and in fact the holding of risky (nonmonetary) assets increases with wealth. This would tend to rule out the quadratic function. However, intuition and the evidence fail with respect to relative risk aversion and provide no basis for discriminating among the other functions. The formulation and data analysis for each of these functions follows.

The hypothesis that utility is a quadratic function of the investor's wealth is used extensively in the literature. In order to meet Tobin's conditions that the function be monotonically increasing and strictly concave we must restrict the quadratic to the rising portion of the curve. This is done by setting the condition  $0 > \alpha > (-1/W)$ . With future wealth uncertain the expected value of utility is

$$(2) \quad E(U) = \sum p_j W_j + \frac{\alpha}{2} \sum p_j W_j^2$$

where  $p_j$  is the probability that wealth is  $W_j$ , and aversion to risk implies  $\alpha < 0$ . For any of the profitable gambles in Aipotu

<sup>3</sup> It is well known that the *log* functions are just specific forms of the power functions when the  $\delta$  value is equal to zero, and these five specific forms collapse to three general functions. However, as the finance and decision theory literature to date appears to treat them distinctly we shall continue with this convention.

$$\begin{aligned}
 E(U) &= \sum p_j [W + G(R_j - 1)] \\
 &+ \frac{\alpha}{2} \sum p_j [W + G(R_j - 1)]^2 \\
 (3) \quad &= [W + G(\bar{R} - 1)] \\
 &+ \frac{\alpha}{2} [W^2 + 2WG(\bar{R} - 1) + G^2V]
 \end{aligned}$$

Here,  $W$  = wealth before the play decision,  $G$  = amount invested in the gamble,  $\bar{R}$  = expected value of the gamble's payoff, and  $V$  = variance of the payoff per dollar invested. Taking the derivative of  $E(U)$  with respect to  $G$  and setting it equal to zero, we find that  $E(U)$  is maximized by the investment

$$\begin{aligned}
 G^* &= \frac{(\bar{R} - 1)(1 + \alpha W)}{-\alpha V} \\
 (4) \quad &= \frac{\bar{R} - 1}{-\alpha V} - \frac{\bar{R} - 1}{V} W
 \end{aligned}$$

Since  $\alpha < 0$ ,  $G^*$  increases with  $\bar{R} - 1$ , the gamble's rate of return, and it decreases with  $V$ , its variance. The optimal play  $G^*$  varies among individuals with their values for  $\alpha$ , and  $G^*$  varies inversely with wealth.

If investor behavior is described by a utility function that is quadratic in the rate of return on the investor's wealth

$$(5) \quad E(U) = \sum p_j r_j + \frac{\alpha}{2} \sum p_j r_j^2$$

where  $p_j$  is the probability that the rate of return on the investor's wealth will be  $r_j$ . For any one of the profitable gambles in Aipotu

$$\begin{aligned}
 E(U) &= \sum p_j [G(R_j - 1)/W] \\
 &+ \frac{\alpha}{2} \sum p_j [G(R_j - 1)/W]^2 \\
 (6) \quad &= G(\bar{R} - 1)/W + \frac{\alpha}{2} [G^2V/W^2]
 \end{aligned}$$

Proceeding as before  $E(U)$  is now maximized by the investment

$$(7) \quad G^* = -\alpha \frac{\bar{R} - 1}{V} W$$

Given the parameters of the gamble,  $\bar{R} - 1$  and  $V$ , the optimum play is proportional to wealth. The proportionality factor  $-\alpha$  varies among investors with their personality.

On the assumption that all participants in our experimental game have the same value for  $\alpha$ , we can test both of the above hypotheses by running the regression

$$(8) \quad (G/Q) = \beta_0 + \beta_1(W - C)$$

$G$  is the amount invested,  $Q = (\bar{R} - 1)/V$  for the gamble selected,<sup>4</sup> and  $W - C$  is the individual's wealth net of his consumption at the time he makes his investment decision. If investor utility functions are quadratic in wealth, we would expect to find  $\beta_0 = 1/(-\alpha_0) > 0$ , and  $\beta_1 = -1$ . If investor utility functions are quadratic in the rate of return on wealth, we would expect to find  $\beta_0 = 0$  and  $\beta_1 > 0$ . If  $\alpha$  varies among individuals, under a quadratic wealth function,  $\beta_0$  varies among individuals and  $\beta_1 = -1$  remains true. Under a quadratic rate of return function  $\beta_1$  varies among individuals and  $\beta_0 = 0$  remains true.

The parameter estimates of equation (8) obtained for the individuals described in Section II were

$$(9) \quad (G/Q) = 83,210 + .692(W - C) \\
 (8,365) \quad (.049)$$

The numbers in parentheses are the standard errors of the parameter estimates and the coefficient of determination was .36. It is clear that the parameter estimates are not consistent with any of the utility functions described above. The quadratic wealth function requires  $\beta_1 < 0$  and our

<sup>4</sup> Dividing each observation on  $G$  by  $(\bar{R} - 1)/V$  for the gamble selected makes the four gambles equivalent.

estimate is .692 with a standard error of .049. The quadratic rate of return function requires that  $\beta_0 \sim 0$  and it is equal to \$83,210 with a standard error of \$8,365.

The failure of the data to support either of the quadratic utility functions may be due to differences in  $\alpha$  among the participants that were suppressed by equation (9). To establish whether differences in  $\alpha$  among the participants were responsible for our failure to obtain  $\beta_1 = -1$ , the condition for the quadratic wealth function to hold, we ran the regression

$$(10) \quad (G/Q) = \sum_{j=2}^{34} \beta_{0j} X_j + \beta_1 (W - C)$$

where the dummy variable  $X_j = 1$  if  $G/Q$  is for the  $j$ th individual and zero otherwise. The constant term for the  $j$ th individual, therefore, is  $\beta_{0j}$ , and it is an estimate of  $-1/\alpha$  for the  $j$ th individual on the assumption that he has a utility function that is quadratic in wealth. The dummy variables raised the coefficient of determination from .36 to .62. All the constant terms were positive with two below \$50,000 and five above \$150,000, and all but three of the standard deviations of  $\beta_{0j}$  fell between \$20,000 and \$25,000. However,  $\beta_1 = .436$  with a standard deviation of .062. Hence, the data do not support the hypothesis that a quadratic in wealth utility function describes investor behavior.

The quadratic rate of return utility function implies that  $\beta_0 = 0$  for all individuals, and  $\beta_1 = -\alpha > 0$  varies among them with their aversion to risk. Accordingly, we ran the regression

$$(11) \quad G/Q = \beta_0 + \beta_1 (W - C) + \sum_{j=2}^{34} \psi_j Z_j$$

where the dummy variable  $Z_j = W - C$  if  $G/Q$  is for the  $j$ th individual and zero otherwise. The coefficient of  $W - C$  for the  $j$ th individual,  $j = 2 \rightarrow 34$  is  $\beta_{1j} = \beta_1 + \psi_j$ . The  $\beta_{1j}$  fluctuated over a fairly wide range.

Two were greater than one, two were less than zero, and another eleven fell outside the interval .20 to .60. However,  $\beta_0 = \$109,500$  with a standard error of \$9,710, and the data do not support the hypothesis that the relation between investment and wealth is explained by a quadratic rate of return utility function.

The above analysis of the data provide no basis for rejecting the hypothesis that some of the individuals are described by one or the other of the above utility functions. We ran equation (8) for each of the thirty-four individuals, and for some of them the parameter estimates were consistent with either of the above utility functions. However, due to the small number of observations for each individual, the limited range of variation in  $W - C$  for some individuals and the frequency with which individuals did not change their play in response to short-run variation in  $W - C$ , the parameter estimates varied over a wide range and had high standard deviations. It therefore was difficult to draw any conclusions from these statistics.

Under the famous Bernoulli logarithmic function

$$(12) \quad E(U) = \sum p_j \ln W_j$$

Proceeding as before, the play on any of the four gambles that maximizes  $E(U)$  is

$$(13) \quad G^* = \frac{\bar{R} - 1}{-(R_1 - 1)(R_2 - 1)} W$$

Here  $R_1$  and  $R_2$  are the two possible pay-offs on a gamble. The Bernoulli function differs from the quadratic rate of return function in that  $G^*/W$  is independent of the investor's personality.

The third function commonly referred to is the power function of the form:

$$(14) \quad E(U) = \sum p_j W_j^\delta$$

where for risk aversion  $\delta$  is constrained by

$$(15) \quad G^* = \frac{1 - \left[ \frac{p_b(1 - R_b)}{p_r(R_r - 1)} \right]^{1/(1-\delta)}}{(1 - R_b) + (R_r - 1) \left[ \frac{p_b(1 - R_b)}{p_r(R_r - 1)} \right]^{1/(1-\delta)}} \cdot W$$

$0 < \delta < 1$ . As above, it can be shown that the optimizing strategy is defined by equation (15). By setting the  $\delta$  value of (15) equal to zero this expression reduces to (13) and it is obvious that the logarithmic function is simply a limiting case of the power function where  $\delta = 0$ .

Close relatives of the power and *log* functions are the adjusted power and *log* functions whose characteristics seem quite attractive when comparing the Pratt descriptions of Table 6 to the statistics of Table 5. In Table 5 it is evident that  $G$  does increase with wealth and that  $G/(W - C)$  tends to decrease with wealth just as this function would indicate. The optimal values of  $G^*$  under these adjusted functions are the values in equations (13) and (15) with  $W$  replaced by  $W + k$ .

Consequently, we can test all four hypotheses by running a regression of the form given in equation (16).

$$(16) \quad (G/Q) = \beta_0 + \beta_1(W - C)$$

In equation (16)  $Q$  is the ratio  $G^*/W$  given by equation (15), and  $\delta$  takes on different values in the range  $0 \leq \delta < 1$ . Each assignment of a value to  $\delta$  is a different hypothesis and involves another regression.

The conditions for the logarithmic or power functions to be true are that  $\beta_0 = 0$  and  $\beta_1 = 1$  with the appropriate value of  $\delta$  in equation (16). The condition for any of the adjusted utility functions (with  $W + k$  replacing  $W$  in equations (13) and (15)) to be true is only that  $\beta_1 = 1$ . Table 7 presents the regression results for equation (16) with  $\delta = 0, .3, .6$ , and  $.9$ . It is clear that all the  $\beta_1$  coefficients are significantly less than one, and none of the utility functions

accurately describe the behavior of the participants in the experimental game. Also the  $t$ -values for  $\beta_0$  indicate it is significantly greater than zero.

TABLE 7—REGRESSION RESULTS FOR THE HYPOTHESIS THAT UTILITY IS A LOGARITHMIC OR A POWER FUNCTION OF WEALTH

$\delta$	$\beta_0$	$t(\beta_0=0)$	$\beta_1$	$t(\beta_1=1)$	$R^2$
$\delta=0$	79,185	18.3	.616	-8.43	.41
$\delta=.3$	55,173	18.3	.489	-18.0	.41
$\delta=.6$	32,110	18.3	.250	-40.6	.41
$\delta=.9$	15,348	17.6	.122	-95.9	.40

#### IV. An Explanation of the Behavior

The limitations of each of the utility functions examined in explaining the data persuaded us to look further. Latané and Tuttle have proposed the following solution to the optimal portfolio policy problem. Let  $p_j$  be the probability that a gamble will pay  $R_j$ . If the fraction  $\lambda$  of wealth is put in the gamble, and the fraction  $1 - \lambda$  is loaned at a zero interest rate, the geometric mean rate of return on or growth in wealth is

$$(17) \quad \theta = \prod_{j=1}^n (1 - \lambda + \lambda R_j)^{p_j} - 1$$

Latané and Tuttle have shown that if an individual repeatedly puts the fraction  $\lambda$  of his wealth in the gamble, as the number of trials become large "... the *ex post* compound average return approaches the *ex ante* geometric mean return ..." (p. 361). In other words, an individual can be practically certain that after a large number of trials his wealth will have grown at the compound rate  $\theta$ . They concluded, quite plausibly, that regardless of the in-

dividuals utility function, maximizing the geometric mean rate of growth in wealth dominates any other policy when repeated investment over a large number of periods is possible.<sup>5</sup>

The value of  $\lambda$  that maximizes  $\theta$  for each of our four profitable gambles maximized  $E(U)$  under a logarithmic utility function. From equation (13) we see that  $\theta$  is maximized by  $\lambda^* = (\bar{R}-1)/-(R_1-1)(R_2-1)$ . For gamble 2,  $\lambda^* = .667$ , and looking back at Table 5, we see that while  $\lambda = G/(W-C)$  is initially much larger than  $\lambda^*$ , it asymptotically falls towards a value slightly below  $\lambda^* = .667$  as  $W-C$  rises above \$150,000. Table 8 reveals this is true regardless of the gamble selected. In Table 8,  $Z$  is the ratio of the fraction of wealth played on a gamble to the Bernoulli-Latané optimal fraction. That is

$$(18) \quad Z = \left[ \frac{G}{W-C} \right] / \left[ \frac{\bar{R}-1}{-(R_1-1)(R_2-1)} \right]$$

When  $Z=1$  the fraction of wealth played is the Bernoulli-Latané optimum. As  $W-C$  rises from the \$150,000-\$175,000 bracket to \$350,000 and over,  $Z$  falls gradually from 1.12 to .85. On the other hand, as  $W-C$  falls from the \$150,000-\$175,000 bracket to less than \$25,000,  $Z$  rises rapidly to 3.48 which is close to the legal limit on the scale of play.

To understand this behavior, let us first consider the profitability of the gambles. The arithmetic mean rate of return on wealth for an individual who put all his wealth on gamble 2 is 10 percent, which seems adequate for a wealth of \$100,000 or more and a minimum consumption of \$5,000. However, it may be surprising to learn that the geometric mean rate is only

TABLE 8—VARIATION IN SCALE OF PLAY WITH WEALTH MINUS CONSUMPTION

W-C Class	Frequency	Mean W-C	Scale of Play Z
Less than 25,000	19	\$15,521	3.48
25,000- 50,000	23	38,817	2.82
50,000- 75,000	26	62,800	2.02
75,000-100,000	39	89,456	1.69
100,000-125,000	62	112,834	1.35
125,000-150,000	44	135,545	1.28
150,000-175,000	56	162,075	1.12
175,000-200,000	16	185,187	.97
200,000-225,000	19	213,032	1.02
225,000-250,000	19	235,384	1.17
250,000-275,000	7	261,943	.87
275,000-300,000	8	286,125	.98
300,000-325,000	7	308,000	.79
325,000-350,000	5	336,520	.70
350,000 and over	11	442,636	.85

2.5 percent and under the Bernoulli optimal play,  $\lambda^* = .667$ , the geometric mean rate of return is raised to  $\theta^* = .033$ . No play on any gamble provides a higher geometric mean rate of return than 3.3 percent. Therefore, the maximum possible long-run compound average rate of growth in wealth with no consumption is only 3.3 percent.

An individual with a wealth of \$150,000 or more might reasonably plan on limiting the fraction of his wealth consumed to  $c = \theta^*$ . Consequently, he might reasonably expect to survive in the long run under a Bernoulli-Latané investment strategy, and Table 8 indicates that  $Z \sim 1$  when  $W-C > \$150,000$ .

However, an individual whose wealth is below \$150,000, certainly if it is below \$100,000, is faced with a difficult choice. With \$5,000 the minimum level of consumption and that figure providing an unattractive standard of living, he cannot find a consumption investment strategy that satisfies  $c < \theta^* = .033$ , and he has the following alternatives. One is to set  $Z \sim 1$  and be practically certain that he will be wiped out in the long run. The other is to set  $Z > 1$  with the result that the probabil-

<sup>5</sup> Hakasson (1969) provides additional theoretical support for maximizing the geometric mean rate of return.

ity of being wiped out in the short run becomes very large, but there is also a rise in the probability that his wealth in the short run will rise to a level consistent with long-run survival at a reasonable standard of living.

To illustrate, consider an individual with  $W = \$50,000$ . With  $C = \$5,000$ , the optimal play on gamble 2,  $\lambda^* = .667$  will raise his wealth to  $\$60,000$  if the outcome is favorable. This provides no perceptible improvement in his situation, while an unfavorable outcome results in  $W = \$37,000$ , a marked deterioration in his situation. On the other hand, the probability is .25 that he can win on two successive trials. If he plays  $\$145,000$  on the first trial, wins, plays  $\$145,000$  on the second trial, and wins again, his wealth rises to  $\$195,000$ . Hence, there is a probability of .25 that a bold strategy will provide him with a viable level of wealth. Of course, there is a probability of .5 that he will be wiped out on the first trial. There is also a probability of .25 that he will win and then lose and be somewhat better off at the end of two periods with  $W = \$79,000$ .

Regardless of what one thinks of the soundness of this strategy, the data indicates it is the policy followed by the participants. That is, as wealth fell below the level necessary to provide a reasonable expectation of long-run survival at an adequate standard of living, the participants raised the scale of play above the Bernoulli-Latané optimum. Doing so increased both the probabilities of immediate ruin and of long-run survival. They preferred this state of affairs to small probabilities of short-run ruin and of long-run survival.

## V. Conclusions

The relation between investment and wealth for the subjects of our experimental game may be summarized as follows. All individuals when reduced to circumstances analogous to being on welfare and only

under those circumstances showed a preference for risk. They played what they could on a lottery—a highly skewed unfair gamble. Under ordinary circumstances they had aversion to risk in that they invested in profitable gambles and limited the amount invested in relation to wealth. More specifically they exhibited decreasing absolute and increasing relative risk aversion in that the amount invested rose while the proportion invested fell as wealth increased. This summary statement, however, does not fully explain the participant's behavior.

Since the circumstances of the game may have had a significant influence on behavior, it is advisable to recall them. First, each participant was required to maintain a minimum consumption level as long as his wealth allowed it. This is considered a realistic replication of a world where individuals with no means of support are provided a subsistence standard of living and where proprietors have the alternative of taking employment with someone else. Second, the participants were informed as to the standard of living provided by various consumption levels and advised to make realistic consumption decisions. Under these circumstances consumption rose absolutely but fell relatively as wealth increased, at least up to some level of wealth. The parameters of the relation, it would seem, were influenced both by the consumption levels considered minimum and comfortable, and by the profitability of the available investment opportunities. The latter in a real sense determine an individual's wealth.

At levels of wealth above  $\$150,000$ , investment approximated the fraction of wealth that maximizes the average long-run compound rate of return on wealth—a Bernoulli-Latané policy, and consumption was a slightly larger fraction of wealth. Both fractions of wealth fell slightly as wealth increased. As wealth fell towards

zero from the \$150,000 level both the fraction of wealth consumed and the fraction of wealth invested rose at increasing rates. The latter not only increased the individual's risk, but it also reduced the geometric mean or long-run rate of return on wealth. The explanation of this behavior lies in the fact that as wealth fell below \$150,000 the rise in the fraction of income consumed made it increasingly difficult to adopt an investment strategy which provided a reasonable expectation of survival in the long run. Hence, as long-run prospects deteriorated under a Bernoulli-Latané policy the participants become increasingly willing to risk ruin in the short run in order to increase the probability that good luck in the short run will create a viable situation in the long run. The conclusion this suggests is that an individual's attitude towards risk is materially influenced by his circumstances.

Information on behavior obtained from experimental games is always suspect. However, the great difficulties involved in obtaining meaningful data on real portfolios and the irrelevance of data obtained from gambles which involve real but trivial stakes severely limit those lines of inquiry for obtaining information on the relation between investment and wealth. That plus the reasonableness of the results obtained argue for further experimentation with portfolio games of the type described here. In addition to replicating the experiment with other groups, it would seem desirable to vary it in a number of ways. One would be to change the profitability and risk of the investment opportunities, including perhaps paying interest on wealth not invested in the risk asset. Another would be to provide the participants with levels of wealth that are very large in relation to middle class standards of living, so that security by such standards would not be a serious consideration. Finally, the indi-

vidual might be given both an initial amount of wealth and periodic income from employment.

#### REFERENCES

- M. Freimer and M. J. Gordon, "Investment Behavior with Utility a Concave Function of Wealth," in K. Borch and J. Mossin, eds. *Risk and Uncertainty*, New York 1968.
- N. H. Hakansson, "Optimal Investment and Consumption Strategies for a Class of Utility Functions," Working Paper No. 101, Western Management Science Institute, 1966.
- , "Capital Growth and the Mean-Variance Approach to Portfolio Selection," Working Paper No. 277, Univ. California, Berkeley 1969.
- H. A. Latané, "Criteria for Choice among Risky Ventures," *J. Polit. Econ.*, Apr. 1959, 67, 144-55.
- and D. L. Tuttle, "Criteria for Portfolio Building," *J. Finance*, Sept. 1967, 23, 359-73.
- J. Lintner, "Security Prices, Risk and Maximal Gains from Diversification," *J. Finance*, Dec. 1965, 20, 587-616.
- H. M. Markowitz, "*Portfolio Selection: Efficient Diversification of Investments*," Cowles Foundation Monograph, No. 16, New York 1959.
- J. Mossin, "Equilibrium in a Capital Asset Market," *Econometrica*, Oct. 1966, 34, 768-83.
- E. H. Neave, "Multi-Period Consumption Investment Decisions and Risk Preference," unpublished paper, Northwestern Univ.
- J. W. Pratt, "Risk Aversion in the Small and the Large," *Econometrica*, Jan./Apr. 1964, 32, 122-36.
- W. F. Sharpe, "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk," *J. Finance*, Sept. 1964, 19, 425-42.
- J. Tobin, "Liquidity Preference as Behavior Towards Risk," *Rev. Econ. Stud.*, Feb. 1958, 25, 65-85.
- M. R. Yaari, "Convexity in the Theory of Choice under Risk," *Quart. J. Econ.*, May 1965, 79, 278-90.

# Social Returns to Public Information Services: Statistical Reporting of U.S. Farm Commodities

By YUJIRO HAYAMI AND WILLIS PETERSON\*

An important function of government is the collection and reporting of information useful for decision making in both the public and private sectors. In this study, we attempt to develop a theoretical framework for estimating the social returns to government expenditures on such public information services.

As an illustration of its possible use, the technique is applied to information reported by the Statistical Reporting Service of the U.S. Department of Agriculture (*USDA*). In this example, we attempt to measure the marginal social returns of reducing the sampling error of crop and livestock statistics reported by the *USDA*. Although the purpose of statistical reporting is to facilitate decision making in both the public and the private sectors, our methodology applies only to the private sector. Since our estimates of social returns do not include the gains due to a better resource allocation by public agencies, these estimates should represent the lower bounds of the social returns.

## I. Theoretical Framework for Estimating Social Returns

In this section we attempt to develop the theory and method of estimating the social

returns to statistical reporting. Alfred Marshall's social welfare and social cost concepts provide the basic theoretical framework.<sup>1</sup> Social welfare is defined as the area under the demand schedule; and social cost, or opportunity cost, is defined as the area under the supply schedule.

Assuming rational profit and utility maximizing behavior by producers, marketing firms and consumers, a sampling error in statistical reporting of the production or the stock of commodities can be expected to lead to a net decrease in social welfare. Erroneous information causes producers to make erroneous production decisions and also distorts optimal inventory carryovers. Hence, marginal improvements in the accuracy of these statistics reduce the social cost of misinformation, which in turn can be considered as an increase in net social welfare. By relating the marginal improvements in the net social welfare to the marginal cost of providing more accurate information, we can estimate marginal social benefit-cost ratios for the various levels of accuracy of the information.

We have developed two models for estimating the social returns to the improvements in information: (a) an inventory

\* Tokyo Metropolitan University and University of Minnesota, respectively. We are indebted to Harry Trelogan, Statistical Reporting Service, U.S. Department of Agriculture for stimulating our interest in this problem and to W. E. Kibler for providing pertinent data and information concerning the cost of the sample survey for the statistical reporting service. We also wish to thank George Borts, K. E. Egertson, J. P. Houck, Mathew Shane, T. W. Schultz, and an anonymous

referee for constructive comments on a previous draft of this paper. Of course, they are not responsible for possible errors which may remain.

<sup>1</sup> See Marshall, pp. 124-133, 140, and 810-812. Our approach is along the tradition of public goods. For a classical theoretical study, see Harold Hotelling. For an empirical study, see Zvi Griliches.

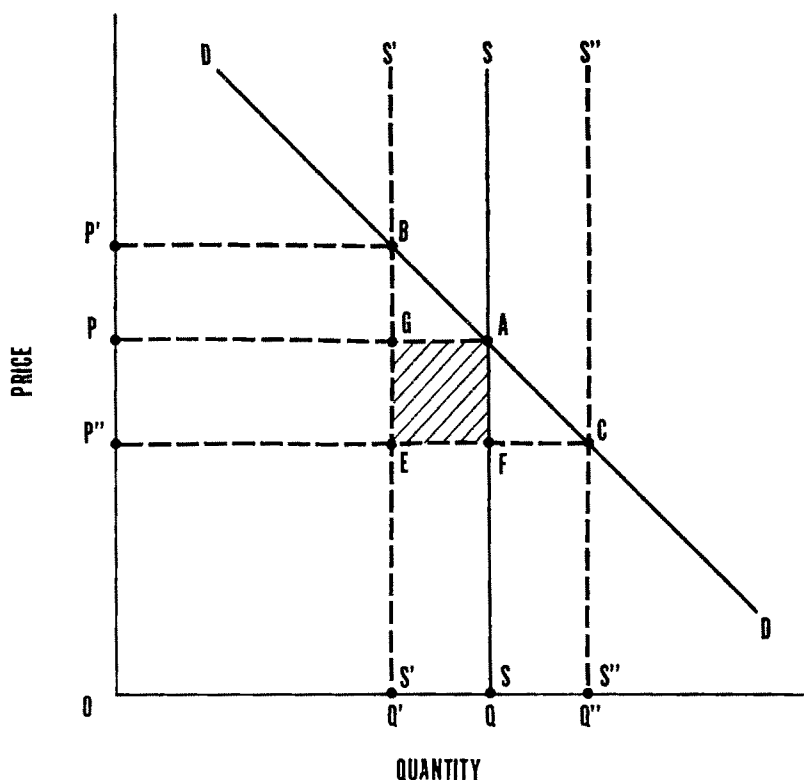


FIGURE 1—INVENTORY ADJUSTMENT MODEL

adjustment model and (b) a production adjustment model.

#### *Inventory Adjustment Model*

The inventory adjustment model applies to situations where production cannot be altered significantly in response to output predictions, but where there is an opportunity for inventory holders to adjust stocks. A good example occurs in agriculture in the case of food and feed grains. Once the crops are planted, it is usually not profitable for producers to significantly expand or contract the output. On the other hand, it is relatively easy and inexpensive to store these commodities. In this case any market supply adjustment is possible mainly through adjustment in inventories.

For products of this type, the social cost

of misreporting of future production, through such errors as acreage or yield estimates, arises because of distortions in the optimum consumption patterns of the products. Because products of this type are produced during a relatively short period of time within the year, their consumption patterns depend very much on the inventory policy of marketing firms. For example, the expectation of an abnormally small crop in the forthcoming production period and of a higher price can be expected to result in a decreased rate of inventory depletion during the remainder of the current period. This in turn results in increased prices and a decreased rate of consumption during the current period.

This situation is illustrated in Figure 1. We assume in this case that production response to a price change can be approxi-

mated as being perfectly inelastic during the production period, as denoted by the supply curve  $SS$ . The market demand schedule for the commodity is denoted by  $DD$ .

Suppose the statistical reporting agency estimates the current period production as  $OQ'$  as opposed to the actual or "true" production  $OQ$ . Inventory holders, in forming price expectations for the coming period, expect the average price to equal  $OP'$ . In other words, they would expect the future price to be higher by  $PP'$  (or  $BG$ ) than would be the case had no error been involved in the production estimate. Consequently inventory holders find it profitable to decrease their rate of inventory depletion for the remainder of the year, until current price has risen by  $PP'$ . Consumption then would contract to  $OQ'$ , or by the amount  $Q'Q$ . In turn, the inventory carry-over into the next production period would be increased by the same amount,  $Q'Q$ . As a consequence, the reduction in consumption during the current period would reduce consumer welfare by the area  $ABQ'Q$ .

Because of the abnormally large carry-over into the next period, we assume that the next period supply would increase by the amount  $Q'Q$  which is equal to  $QQ''$  in Figure 1. Hence the total quantity placed on the market during the next period would be the "true" production  $OQ$  plus the increased carry-over  $QQ''$ . The result would be a decrease in the average price down to  $OP''$  as opposed to price  $OP$  which would have prevailed had there been no reporting errors. The decrease in price, however, results in an increase in consumption during the next period by the amount  $QQ''$ . Thus total consumer welfare is increased during the next period by  $ACQ''Q$ . The overall result of reporting errors that gave rise to the decline in current consumption and the increase in future consumption is a net loss in consumer welfare

equal to rectangle  $AGEF$  (area  $ABQ'Q$  minus area  $ACQ''Q$ ), the shaded area in Figure 1, assuming that the demand curve is linear.

The same amount of net welfare loss would have resulted from an erroneous overestimate of production, that is, if  $OQ''$  would have been predicted instead of  $OQ'$ . Since the errors in statistical reporting (mainly due to sample errors in the example presented in a later section of this paper) can be expected to be random, inventory costs can be expected to average out to zero over a period of years.

Assuming a linear demand curve, the area of rectangle  $AGEF$ , which is  $AG \cdot AF$ , can be estimated if we have an estimate of the price elasticity of demand ( $\alpha$ ) of the commodity. Since  $AF$  is found by multiplying the error in production reporting  $QQ''$  (or  $QQ'$ ) by the absolute value of the slope of the demand curve  $(1/\alpha)(p/q)$ , we obtain<sup>2</sup>

$$\text{area } AGEF = \epsilon^2 p q \frac{1}{\alpha}$$

where  $q$  is the true quantity of production ( $OQ$ );  $p$  is the equilibrium price ( $OP$ ); and  $\epsilon$  is the error in quantity of production reported as a proportion of the true production ( $QQ' = QQ'' = \epsilon q$ ).

#### *Production Adjustment Model*

Next let us consider the situation where producers have an opportunity to adjust output in response to additional information, as illustrated by the upward-sloping supply schedule ( $SS$ ) in Figure 2. In the context of the example to be presented in a later section of this paper, those commodities for which a continuous adjustment in production is possible include mainly livestock products.

<sup>2</sup> In this case  $\alpha$  is the absolute value of the price elasticity of demand.

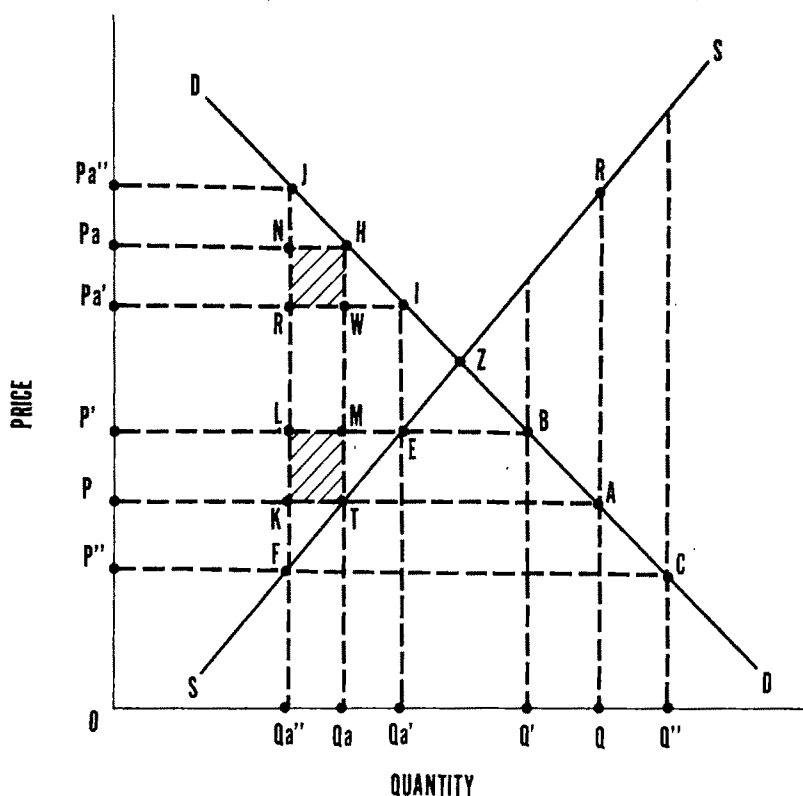


FIGURE 2—PRODUCTION ADJUSTMENT MODEL

A basic assumption of the production adjustment model is that producers adjust output along their supply schedules in response to changes in their price expectations. Furthermore, it is assumed that changes in price expectations come about as a result of new information on expected output provided by statistical reporting agencies. This implies a process of adjustment similar to that of the well-known cobweb model. As a result certain stability conditions became important. We will consider these conditions in more detail in the following section. First let us develop the model.

Suppose, to begin, that producers are unaware that their actual production in the forthcoming period would equal  $OQ$  if current production plans materialize. If a sample survey of production could ac-

curately predict  $OQ$  and producers have information on the nature of demand for the commodity, the predicted price in the coming period would be  $OP$ . Reacting to price  $OP$ , producers cut back output and actually place quantity  $OQ_a$  on the market, which results in price  $OP_a$  in the coming period.

Of course, the below equilibrium output involves a misallocation of resources and therefore leads to a social loss to society. Assuming perfect competition, with no externalities, the marginal cost of output  $OQ_a$  is  $OP$ , as shown by the supply schedule. However, at quantity  $OQ_a$  we see that price  $P_a$  will prevail, indicating that society values the marginal unit of this product more highly than it values the products given up to produce it. At the margin this difference is equal to  $HT$  on

the price axis. Adding to the output of this product would continue to add to net social welfare by progressively smaller amounts, until the equilibrium quantity is reached. Hence, the total net social loss of producing  $Q_a$  instead of the equilibrium quantity is equal to triangle  $ZHT$  in Figure 2.

Suppose, however, that the statistical reporting agency overestimates  $OQ$  and predicts  $OQ''$  instead, because of a sampling error. On the basis of this information, producers expect price  $OP''$  and react by actually producing  $OQ_a''$ . In this case, the net social loss would increase to triangle  $ZJF$ . Let us assume however, that sampling errors can be expected to occur at random, and that there is a 0.5 probability that the statistical reporting agency will underestimate production and predict  $OQ'$ . Now the net social loss would decrease (from the initial situation) to triangle  $ZIE$ .

It is important to recognize, however, that the expected value of the reduction in social loss due to an underestimate of production will not offset the increase in social loss due to an overestimate. As shown in Figure 2, the overestimate of output by  $QQ''$  results in an addition to social loss by area  $HJFT$ . But an underestimate of production reduces social loss by area  $IHTE$ .<sup>3</sup> The difference between these two areas is equal to the two shaded rectangles in Figure 2, area  $TMLK$  plus area  $HNRW$ . If the probability of either an overestimation or an underestimation is 0.5, the expected value of the net addition to social loss due to the random sampling error, in any given year, is

$$\frac{1}{2} (\text{area } TMLK + \text{area } HNRW)$$

<sup>3</sup> This does not imply that the statistical reporting agency can always reduce social loss by biasing their estimates on the low side. If actual output happened to be less than equilibrium output, the under-estimate error increases social loss instead of reducing it.

Assuming linear demand and supply curves, the areas of the rectangles  $TMLK$  and  $HNRW$  can be calculated if we have estimates of the price elasticity of demand ( $\alpha$ ) and the price elasticity of supply ( $\beta$ ) of the commodity in question. The area of rectangle  $TMLK$  is equal to  $TM \cdot TK$ . Since  $TM$  is found by multiplying the error in reporting,  $QQ' (= QQ'')$ , by the absolute value of the slope of the demand curve  $(1/\alpha)(p/q)$  and  $TK$  is found by multiplying  $TM$  by the inverse of the slope of supply curve  $(\beta q/p)$ , we obtain

$$\text{area } TMLK = \epsilon^2 pq \frac{\beta}{\alpha^2}$$

where  $q$  is the quantity of true production;  $p$  is the corresponding price on the demand schedule; and  $\epsilon$  is the quantity error in statistical reporting as a proportion of the true quantity ( $QQ' = QQ'' = \epsilon q$ ).

The area of rectangle  $HNRW$  is equal to  $HN \cdot HW$ . The distance  $HN$  is equal to  $TK$  while  $HW$  is found by multiplying  $HN$  by the slope of the demand curve  $(1/\alpha)(p/q)$ . Thus, we obtain

$$\text{area } HNRW = \epsilon^2 pq \frac{\beta^2}{\alpha^3}$$

Consequently, the net social cost due to the error in production reporting (sampling error) is given by

$$\begin{aligned} & \frac{1}{2} (\text{area } TMLK + \text{area } HNRW) \\ &= \frac{1}{2} \epsilon^2 pq \left( \frac{\beta}{\alpha^2} + \frac{\beta^2}{\alpha^3} \right) \end{aligned}$$

The above formulation applies equally to the case where the actual or true production is smaller than the equilibrium production. However, now the social loss occurs because society gives up other goods and services that it values more highly than the commodity in question. Assuming linear demand and supply

curves, the magnitude of net social loss due to an error in statistical reporting is the same regardless of whether the actual production is larger or smaller than the equilibrium output.

As a special case, actual output can coincide with the equilibrium quantity. But a statistical reporting error still results in a net social loss in this case; the same formula we derived,

$$\frac{1}{2} \epsilon^2 pq \left( \frac{\beta}{\alpha^2} + \frac{\beta^2}{\alpha^3} \right),$$

can be used to estimate this loss.

#### *Stability Conditions<sup>4</sup>*

Because the process of adjustment implied by the production adjustment model is of a quasi-cobweb nature, it is important that we investigate the stability of the model. Recall that the cobweb model converges only if the supply curve is steeper (less elastic) than the demand curve at least in the vicinity of the equilibrium point.

If the production adjustment model correctly describes the process of adjustment that occurs in response to output and price information supplied by a statistical reporting agency, then a positive social return is obtained from this information only if the cobweb is stable. This can be seen from Figure 2. Suppose for the sake of argument that the statistical reporting agency does not exist and output  $OQ$  is produced. The social cost of this disequilibrium situation is given by triangle  $ZRA$ . Compare this with the case where a statistical reporting agency is able to predict  $OQ$  with 100 percent accuracy. The production adjustment model implies that producers respond by reducing output to  $OQ_a$  and as a result the social cost is equal to triangle  $ZHT$ . Notice, however, that triangle  $ZHT$ , the social cost with per-

fectly accurate information, is less than triangle  $ZRA$ , the social cost with no information, only if the supply curve is steeper than the demand curve. If the converse is true, then information provided by a statistical reporting agency results in a net loss to society. This is one case where "it pays to be ignorant."

Whether or not society is better or worse off with statistical reporting agencies of the type discussed in conjunction with the production adjustment model appears to depend, therefore, upon whether or not the model is convergent. This, of course, will depend upon the commodity in question. In the case of the *U.S.* agricultural commodities, to which we apply the model in the following section, we have evidence that leads us to believe that the model is convergent. According to estimates by W. A. Cromarty and G. E. Brandow (see Table 2), it appears that for livestock products the demand elasticities are substantially larger than their corresponding supply elasticities. In addition, we do not observe an increasing amount of price instability in the markets for these products particularly in the post-World War II era. In fact, just the opposite appears to be the case, although this may be explained in part at least by a more stable economy during this time. At any rate the evidence clearly supports the hypothesis that the Statistical Reporting Service of the *USDA* provides a positive rather than a negative social return.

We should point out, however, that statistical reporting errors increase social cost irrespective of the market stability condition, and that the formula developed in conjunction with the production adjustment model applies in either case. That is, the two shaded rectangles in Figure 2, denoting the expected increase in net social loss due to random sampling errors, continue to exist regardless of the relative size of triangles  $ZRA$  and  $ZHT$ . Of course, in the unstable case where tri-

<sup>4</sup> We are indebted to George Borts for calling our attention to this problem.

TABLE 1—COSTS OF SAMPLE SURVEY REQUIRED FOR SPECIFIED LEVELS OF TYPICAL SAMPLING ERRORS IN MAJOR U.S. FARM COMMODITIES, AND THEIR CORRESPONDING SAMPLING ERRORS

Area sample	Survey Cost (million dollars)						
	3.40	3.76 <sup>c</sup>	4.13	5.80	7.90	17.10	62.00
Multiple frame sample	3.40	3.76 <sup>c</sup>	4.13	5.60	7.60	13.00	44.20
Typical sampling error in major commodities <sup>a</sup>	3.0	2.5	2.0	1.5	1.0	0.5	0.0
Individual commodities sampling error: <sup>b</sup>	(percent)						
Wheat	3.2	2.6	2.1	1.6	1.1	0.7	0.2
Rye	9.0	7.5	5.9	4.5	3.0	2.0	0.6
Rice	15.8	12.6	9.9	7.8	5.5	3.5	0.8
Corn	2.1	1.8	1.4	1.1	0.8	0.5	0.0
Oats	3.1	2.6	2.1	1.7	1.2	0.7	0.2
Barley	5.4	4.5	3.5	2.7	1.9	1.3	0.5
Potatoes	18.5	15.5	12.6	9.5	6.6	4.2	1.0
Soybeans	3.4	2.8	2.2	1.7	1.2	0.8	0.3
Peanuts	9.5	8.0	6.3	5.0	3.6	2.2	0.8
Tobacco	5.1	4.3	3.4	2.6	1.8	1.2	0.5
Cotton	4.8	4.0	3.1	2.4	1.7	1.1	0.4
Cattle	2.3	1.9	1.3	1.0	0.7	0.5	0.0
Hogs	4.4	3.8	2.9	2.2	1.6	1.0	0.4
Sheep & Lambs	13.1	11.0	8.9	6.8	4.5	3.0	0.7
Poultry	9.2	7.8	6.2	4.8	3.3	2.0	0.5
Eggs	9.2	7.5	5.8	4.5	3.1	1.9	0.6
Milk	5.4	4.5	3.5	2.7	1.9	1.3	0.4

<sup>a</sup> Major commodities refer to items that are produced on most farms in the United States.

<sup>b</sup> Sampling errors in the production characteristics of individual items corresponding to the specified levels of typical sampling error in major U.S. farm commodities.

<sup>c</sup> Linear interpolation.

Source: Data prepared by the Statistical Reporting Service, U.S. Department of Agriculture.

angle ZRA is smaller, the relevant question pertains to the existence of the statistical reporting agency rather than to the size of its sampling errors.

Let us now turn our attention to the measurement of the marginal costs and returns of achieving greater sampling accuracy in the statistical reporting of U.S. agricultural commodities. By comparing these costs and returns we will be able to estimate the marginal benefit-cost ratios to public investment in this activity.

## II. Costs and Returns of Statistical Reporting of Agricultural Production

### *Sample Survey Costs for Alternative Degrees of Accuracy*

For the purpose of reporting and predicting agricultural production, The Statistical Reporting Service of the USDA conducts a nationwide sample survey cover-

ing approximately 150 agricultural commodities. The costs of obtaining specified levels of accuracy in the sample survey are estimated by the Research and Development Branch of the Statistical Reporting Service. These cost estimates for degrees of accuracy ranging from a zero to a 3 percent sampling error for the major farm commodities are presented in Table 1.<sup>5</sup> Also presented in Table 1 are the corresponding sampling errors for each of the individual commodities included in this group.

The Statistical Reporting Service is now shifting its methodology of sampling from an area technique to multiple frame tech-

<sup>5</sup> The data given in Table 1 were developed "on a state-by-state basis and were built up to the national level" (a personal communication with W. E. Kibler, Research and Development Branch, Standards and Research Division, The Statistical Reporting Service, May 8, 1970).

TABLE 2—ESTIMATES OF SOCIAL RETURNS TO REDUCING SAMPLING ERROR, ELEVEN U.S. FARM COMMODITIES, INVENTORY ADJUSTMENT MODEL, 1966-68 AVERAGES<sup>a, b</sup>

	Wheat	Rye	Rice	Corn	Oats	Barley	Pota- toes	Soy- beans	Pea- nuts	To- bacco	Cotton
Price elasticity of demand ( $\alpha$ )	0.02	0.04	0.04	0.03	0.01	0.01	0.1 (million dollars)	0.3	0.2	0.5	0.1
Farm value of production ( $pq$ )	2075	26	464	4882	540	380	598	2534	285	1255	1053
Social loss corresponding to typical sampling error:											
100 $\epsilon$ =3.0 percent	106.2	5.3	289.6	71.8	51.9	110.8	204.7	9.8	12.9	6.5	24.3
2.5	70.1	3.7	184.2	52.7	36.5	77.0	143.7	6.6	9.1	4.6	16.8
2.0	45.8	2.3	113.7	31.9	23.8	46.6	94.9	4.1	5.7	2.9	10.1
1.5	26.6	1.3	70.6	19.7	15.6	27.7	54.0	2.4	3.6	1.7	6.1
1.0	12.6	0.6	35.1	10.4	7.8	13.7	26.0	1.2	1.8	0.8	3.0
0.5	5.1	0.3	14.2	4.1	2.6	6.4	10.5	0.5	0.7	0.4	1.3
0.0	0.4	0.0	0.7	0.0	0.2	1.0	0.6	0.1	0.1	0.1	0.2
Marginal social returns corresponding to:											
100 $\epsilon$ =3.0 to 2.5 percent	36.1	1.6	105.4	19.1	15.4	33.8	61.0	3.2	3.8	1.9	7.5
2.5 to 2.0	24.3	1.4	70.5	20.8	12.7	30.4	48.8	2.5	3.4	1.7	6.7
2.0 to 1.5	19.2	1.0	43.1	12.2	8.2	18.9	40.9	1.7	2.1	1.2	4.0
1.5 to 1.0	14.0	0.7	35.5	9.3	7.8	14.0	28.0	1.2	1.8	0.9	3.1
1.0 to 0.5	7.5	0.3	20.9	6.3	5.2	7.3	15.5	0.7	1.1	0.4	1.7
0.5 to 0.0	4.7	0.3	13.5	4.1	2.4	5.4	9.9	0.4	0.6	0.3	1.1

<sup>a</sup> Social loss =  $\epsilon^2 pq \frac{1}{\alpha}$

<sup>b</sup> Sources, see Table 3.

niques (using lists in conjunction with the area), in order to attain higher accuracy. At present the enumerative and objective yield surveys, using the area technique, are being conducted with a goal of attaining an average sampling error of 2 percent; the cost of these surveys is \$4.13 million. This cost would be similar for the multiple frame sampling scheme down to the 2 percent error level. This 2 percent error objective is based on the fact that the cost of a survey begins to rise rapidly almost with a kink at the 2 percent level of error. For sampling errors of less than 2 percent, the multiple frame technique is more efficient.

A relevant question at this point is whether the marginal cost of attaining greater statistical accuracy represents a socially profitable investment. We can shed some light on this question by comparing the marginal cost of greater accuracy with its accompanying marginal net social return as calculated by the

techniques developed in the previous sections.

#### *Estimation of Marginal Net Social Returns*

In agricultural production it is possible to utilize both the inventory adjustment and the production adjustment models for the various kinds of products. Sampling errors in *crop* reporting data can be evaluated by the inventory adjustment model. In this case, there is little chance to adjust production once the crops have been planted. However, there is ample opportunity for inventory holders to adjust the rate of inventory depletion in response to information on acreages planted and on predicted yields. On the other hand, *live-stock* and livestock products appear to be well suited to the production adjustment model. Here continuous adjustments in production can be made in response to information reported by the government.

TABLE 3—ESTIMATES OF SOCIAL RETURNS TO REDUCING SAMPLING ERROR,  
SIX U.S. FARM COMMODITIES, PRODUCTION ADJUSTMENT MODEL<sup>a</sup>

	Cattle	Hogs	Sheep and lambs	Poultry	Eggs	Milk
Price elasticity of demand ( $\alpha$ )	0.8	0.7	1.8	1.6	0.6	0.6
Price elasticity of supply ( $\beta$ )	0.04	0.1	0.1	0.7	0.3	0.2
			(million dollars)			
Farm value of production ( $pq$ )	8180	4064	246	1758	1981	5745
Social loss corresponding to typical sampling error:						
100 $\epsilon$ =3.0 percent	0.2	1.0	0.7	2.9	11.7	6.2
2.5	0.1	0.7	0.5	2.1	7.0	4.3
2.0	0.0	0.4	0.3	1.3	4.2	2.6
1.5	0.0	0.2	0.2	0.8	2.5	1.5
1.0	0.0	0.1	0.1	0.4	0.9	0.8
0.5	0.0	0.0	0.0	0.1	0.4	0.4
0.0	0.0	0.0	0.0	0.0	0.0	0.0
Marginal social returns corresponding to:						
100 $\epsilon$ =3.0 to 2.5 percent	0.1	0.0	0.2	0.8	4.7	1.9
2.5 to 2.0	0.1	0.0	0.2	0.8	2.8	1.7
2.0 to 1.5	0.0	0.0	0.1	0.5	1.7	0.7
1.5 to 1.0	0.0	0.0	0.1	0.4	1.6	0.5
1.0 to 0.5	0.0	0.0	0.1	0.3	0.5	0.4
0.5 to 0.0	0.0	0.0	0.0	0.1	0.4	0.3

$$^a \text{ Social loss} = \frac{1}{2} \epsilon^2 pq \left( \frac{\beta}{\alpha^2} + \frac{\beta^2}{\alpha^3} \right)$$

Sources: Price elasticity of demand for crops: Commodities except soybeans, tobacco, and cotton from Brandow, p. 59; soybeans from Houck and Mann, p. 20; tobacco from Lyon and Simon, p. 893. Median figure in the estimates of price elasticity for cigarettes; cotton from Donald, Lowenstein, and Simon, p. 61.

Price elasticities of demand for livestock products: Cromarty, p. 572, except sheep and lambs (from Brandow) and eggs (assumed same as in milk).

Price elasticity of supply: Cromarty, p. 573, except sheep and lambs (assumed same as in hogs).

Farm value of production: USDA.

Of course, we might expect some products to have applicability to both models. There are, for example, significant inventories of livestock products in cold storage which could be analyzed by the inventory adjustment model. There are, on the other hand, possibilities of production adjustments in crops, particularly if we consider interregional adjustments. For example, errors in the statistical reporting of the winter wheat acreage in Kansas and Oklahoma may influence decisions to plant spring wheat in Montana and North Dakota. The fact that we apply only one of the adjustment models to each major commodity would seem to imply, therefore, that our estimates of the social re-

turns to improvements in sampling accuracy represent lower bounds of the true returns.

Data for farm value of production ( $pq$ ) were obtained from *Agricultural Statistics*. The time period 1966-68 roughly corresponds to the years for which the costs of the sample survey in Table 1 were estimated. Price elasticities of demand and supply ( $\alpha$  and  $\beta$ ) were collected from various publications (see sources under Table 3).<sup>6</sup>

<sup>6</sup> We attempted to collect the estimates of price elasticities from the studies widely recognized among the profession. It is somewhat difficult to judge the reliability of these elasticities and the possible direction of bias.

TABLE 4—ESTIMATES OF MARGINAL SOCIAL BENEFIT-COST RATIOS CORRESPONDING TO REDUCTION IN TYPICAL SAMPLING ERROR IN THE SURVEY FOR STATISTICAL REPORTING OF FARM COMMODITIES, THE UNITED STATES

Change in typical sampling error	Marginal survey cost <sup>a</sup>		Marginal social returns <sup>b</sup>			Marginal benefit-cost ratio	
	Area sampling (1)	Multiple frame sampling (2)	Inventory adjustment (3)	Production adjustment (4)	Total (5)	(5)/(1)	(5)/(2)
(percent)	(million dollars)						
from 3.0 to 2.5	0.36	0.36	288.8	8.0	296.8	824	824
2.5 to 2.0	0.37	0.37	223.2	5.9	229.1	619	619
2.0 to 1.5	1.67	1.47	152.5	3.6	156.1	93	106
1.5 to 1.0	2.10	2.00	116.3	2.9	119.2	57	60
1.0 to 0.5	9.20	5.40	66.9	1.4	68.3	7.4	13
0.5 to 0.0	44.9	27.1	42.7	0.9	43.6	1.0	1.6

<sup>a</sup> Increases in the cost of sample survey corresponding to changes in the typical sampling error; data from Table 1

<sup>b</sup> Aggregates of marginal social returns; data from Tables 2 and 3.

In Table 2 we present the social losses corresponding to degrees of sampling errors, as opposed to a zero error. The actual  $\epsilon$  values for each commodity are taken from Table 1. For example, in the inventory adjustment model, the  $\epsilon$  value (in percent) for wheat is 3.2 at the 3.0 percent group level sampling error.

The marginal net social returns figures resulting from a reduction in sampling error are presented in Tables 2 and 3. These figures are obtained by subtracting the social cost of a given sampling error from its next higher level. For example, the marginal net social returns for wheat in the inventory adjustment model, because of reducing the typical sampling error from 3.0 to 2.5 percent, is \$36.1 million. This figure is obtained by subtracting the social loss of a 2.5 percent error, \$70.1 million, from the social loss of a 3.0 percent error, \$106.2 million.

### *The Benefit-Cost Ratios*

Based on the estimates of the costs of the sample surveys reported in Table 1 and of the marginal social returns, we calculated the benefit-cost ratios for public investment in increasing accuracy or reducing

sampling error in the survey of agricultural production as being conducted by the Statistical Reporting Service, *USDA*. The results are presented in Table 4.

Marginal costs of the survey for reduction in the typical sampling errors are calculated from the data in Table 1. Marginal social returns corresponding to the reduction in the typical sampling errors are aggregated from the estimates of marginal social returns for individual commodities in Tables 2 and 3.

In spite of the possibility of underestimation of social returns, the benefit-cost ratios calculated by dividing the marginal social returns by marginal social costs are extremely large. For example, our estimates reveal that each extra dollar invested in increasing the accuracy of statistics from the 2.5 to the 2.0 level of error returns more than \$600 worth of benefit to society. And increasing the level of accuracy from 2.0 to 1.5 percent error produces \$90 to \$100 of benefit for each extra dollar invested.

To a certain extent the reliability of our estimates of the marginal social returns and benefit-cost ratios depends on the accuracy of the price elasticities of demand

TABLE 5—MARGINAL SOCIAL BENEFIT-COST RATIOS CORRESPONDING TO AN INCREASE IN THE PRICE ELASTICITIES OF DEMAND ( $\alpha$ ) AND A DECREASE IN THE PRICE ELASTICITIES OF SUPPLY ( $\beta$ )

Changes in typical sampling error	2.5 to 2.0 percent	2.0 to 1.5 percent	
Sample survey method	Area or multiple	Area	Multiple
Marginal benefit-cost ratios using* $\alpha$ 's increased and $\beta$ 's decreased by			
0 percent	619	93	106
10	560	85	96
30	470	71	81
50	434	62	70
100	304	46	52
300	151	23	26
500	103	16	18

\* See original  $\alpha$ 's and  $\beta$ 's in Table 2 and 3.

and supply that we have utilized. Overestimation of the marginal social returns or the benefit-cost ratios would result from either an underestimate of demand elasticities ( $\alpha$ 's), an overestimate of the supply elasticities ( $\beta$ 's), or both. In the interest of obtaining lower bounds to the various benefit-cost ratios we utilized progressively larger demand elasticities and progressively smaller supply elasticities in making our calculations. The results are presented in Table 5.

As expected, the benefit-cost ratios decline using progressively larger  $\alpha$ 's and smaller  $\beta$ 's. However, even when the 500 percent larger  $\alpha$ 's and 500 percent smaller  $\beta$ 's are applied, an extra dollar invested in increasing the accuracy of statistical reporting of the products considered returns over \$100 worth of benefit to society at the 2.5 to 2.0 percent range of accuracy and nearly \$20 of benefit at the 2.0 to 1.5 percent range.

Although the present estimation is very rough and is intended more to illustrate the methodology, it seems apparent that the benefit from the investment in increasing accuracy for agricultural production statistics exceeds its cost by a wide margin. It appears, therefore, that in terms of social welfare maximization criteria it pays

to increase public expenditure to obtain greater accuracy of information concerning agricultural production.

### III. Summary and Conclusions

In this study a methodology was developed to estimate the social returns to investment in the collection and reporting of information. The methodology was applied to the case of reporting of agricultural production statistics by the *USDA*. We found that the social returns exceed the cost of data collection over an extremely wide margin even after adjusting for possible overestimation of the benefit-cost ratios arising from possible errors in the demand and supply elasticities.

In addition to the adjustments just referred to there are a number of other reasons why our results should represent the lower bounds of estimates of the social returns: a) all commodities covered by the same sample survey are not included in the benefit calculations; b) the benefits arising from the better inventory adjustments of livestock products and from the better production adjustments of crops are not included in our calculations; and c) the benefits from the better planning and resource allocations by government agencies are not included in our calcula-

tions. The excess of social benefits over costs would further widen if we were to include these benefits.

Our results suggest that there is an underinvestment in the provision of public information services, at least with respect to statistical reporting in agricultural production. However, this study does not necessarily imply that the government should reduce the output of other public service activities in order to improve information services. The study by Griliches in which the social returns to hybrid corn research were estimated indicates that the benefit-cost ratio for the research is in the order of 70. Peterson's study of poultry research indicates that the ratio is in the order of 20. Thus our results indicate that the social returns to a dollar invested in statistical information service is comparable to the returns in such high pay-off investments as agricultural research.

#### REFERENCES

- G. E. Brandow, *Interrelations Among Demand for Farm Products and Implications for Control of Market Supply*, Agr. Exp. Sta. Bull. 680, Penn. State Univ. 1961.
- W. A. Cromarty, "An Econometric Model for United States Agriculture," *J. Amer. Statist. Ass.*, Sept. 1959, 54, 556-74.
- J. R. Donald, F. Lowenstein, and M. S. Simon, *The Demand for Textile Fibers in the United States*, USDA-ERS Tech. Bull. 1301, Washington 1963.
- Z. Griliches, "Research Costs and Social Returns: Hybrid Corn and Related Innovations," *J. Polit. Econ.*, Oct. 1958, 66, 419-31.
- H. Hotelling, "General Welfare in Relation to Problems of Taxation and of Railway Utility Rates," *Econometrica*, July 1938, 6, 242-69.
- J. P. Houck and J. S. Mann, *An Analysis of Domestic and Foreign Demand for U.S. Soybeans and Soybean Products*, Agr. Exp. Sta. Tech. Bull. 256, Univ. Minn. 1968.
- H. Lyon and J. Simon, "Price Elasticity of Demand for Cigarettes in the United States," *Amer. J. Agr. Econ.*, Nov. 1968, 50, 888-94.
- A. Marshall, *Principles of Economics*, 8th ed., London 1916.
- W. Peterson, "Return to Poultry Research in the United States," *J. Farm Econ.*, Aug. 1967, 49, 656-69.
- U.S. Department of Agriculture, *Agricultural Statistics 1969*, Washington 1969.

# COMMUNICATIONS

## Welfare Economics and Welfare Reform

By GEORGE DALY AND FRED GIERTZ\*

"Often the poor man is not so cold and hungry as he is dirty and ragged and gross. It is partly his taste and not merely his misfortune. If you give him money, he will perhaps buy more rags with it."

Henry David Thoreau in *Walden*

The phrase "welfare system" refers to a collection of institutions which effect transfers of income from one set of individuals to another set whose members are, by some criteria, regarded as less fortunate than the former group. Proposals for changes in the nature of this system have become a topic of deep and continuing political interest in the United States and elsewhere. A central feature of many of these proposals (e.g., the negative income tax) has been the idea that the transfers of income effected by these institutions be in the form of general purchasing power (money) rather than specific commodities.

These proposals for a welfare system emphasizing money transfers rather than transfers "in kind" have gained substantial support from professional economists. Prominent among the reasons for this support has been the notion, grounded in elementary economic theory, that transfers in the form of specific commodities inherently involve a deadweight welfare loss for society. The purpose of this paper is to argue that under a set of highly plausible and well-defined circumstances this notion is incorrect. It is also argued that, from the point of view of practical politics, such proposals for reform may

hurt precisely those individuals whom they are presumably intended to help.

### I. The Nature of the Problem

One of the primary reasons economists prefer to transfer purchasing power is the fundamental proposition of consumer sovereignty: that individuals free to allocate their own incomes will be made better off by their own choices than by the choices of someone else. As such, the economist's preference in this matter is a restatement of his faith in the trading process and the free choice it implies as a vehicle for individual utility maximization and, hence, the collective achievement of Pareto optimal states. When transfers take the form of purchasing power recipients can, by engaging in trade, equate their marginal rates of commodity substitution (*MRS*) to the set of market price ratios (and hence to each other's *MRS*), thereby maximizing their utilities. If they were instead given specific endowments of illiquid commodities of equal market value, they would be unable to engage in trade and there would be no reason to suppose that the particular commodity bundle they were given would be that unique one which would maximize their utility (at that level of income).<sup>1</sup> Thus, recipients can always increase their level of welfare (or,

\* Department of economics, University of Houston and Miami University, Oxford, Ohio, respectively. We would like to express our gratitude to (without implicating) Henry Aaron, James Buchanan, Andrew Gold, and Hajime Hori for helpful comments on an earlier version of this paper.

<sup>1</sup> This argument can be easily visualized in terms of two-dimensional indifference curve analysis. In the context of that analysis, a given cash transfer implies that the recipient is placed on a particular budget line while a specific transfer of commodities of the same value implies a particular point on that budget line. Clearly, this argument applies in a strict sense only if commodities other than money are perfectly illiquid. However, if commodities are less liquid than money and transactions are not costless there will still be a deadweight welfare loss from "in kind" transfers since unnecessary trading costs are imposed upon recipients who must reallocate their budgets subsequent to the receipt of the transfer.

in the limiting case, maintain it) if transfers of a given market value take the form of money rather than illiquid goods. When they are given money rather than goods of the same market value, recipients are made better off, donors are no worse off, and social welfare has therefore increased according to Pareto criteria.

The above argument represents the classical case for purchasing power transfers. Underlying its conclusion is the conception of the maximizing individual used in orthodox price theory—that of an individual who neither affects nor is affected by the welfare of others or, more formally, of an individual with a strictly private utility function. Yet, as has been noted elsewhere, it is impossible to rationalize the existence of many social institutions given such preference functions. (See Kenneth Boulding (1969), William Vickrey, Edgar Olsen (1969).) Indeed, it may be impossible to rationalize many of the one-way transfers of the type which characterize the welfare system without assuming the existence of interdependence between the preference functions of donors and recipients. It becomes necessary to assume that the levels of welfare (however defined) of some individuals exert external effects upon potential donors, effects which cause the donors to wish to affect those levels of welfare through public and private income transfers.<sup>2</sup>

Once one-way transfers are considered in a world of interdependent utility functions the issue of the efficient form of such transfers becomes more complex. Now any transfer must, according to the Pareto criteria, be evaluated on the basis of its effects upon the welfare of *donors* as well as recipients since, by definition, both groups are affected by the welfare of the latter. Depending upon the form taken by the relevant externalities it may also be true that donors are not indifferent to the consumption choices of the recipients. Should this be the case, the welfare superiority of cash transfers is called into serious question.

The remaining sections of this paper are devoted to an examination of the form of

efficient transfers in a world of consumption externalities. To do this a model featuring such externalities is constructed. In the context of this model Pareto optimality and the role of trade in achieving it are examined. Finally, certain policy issues are examined in the light of the theoretical results obtained.

## II. Consumption Externalities, Trade, and Pareto Optimality

### A. The Analytical Framework

Consider a world of perfect knowledge and markets in which there exist  $k+1$  individuals and fixed quantities of  $n$  commodities,  $X_1, \dots, X_n$ . For heuristic purposes it is assumed that  $j=1, 2, \dots, k$  of the individuals have strictly private utility functions. The remaining individual, whom we shall call  $A$ , is subject to external effects from the actions or activities of some but not all of the remaining  $k$  individuals. For purposes of analytical convenience it is assumed that all arguments of all utility functions have first derivatives that are greater than or equal to zero and second derivatives that are less than or equal to zero.<sup>3</sup>

Of crucial importance to the purpose of this paper is the precise form taken by the utility interdependence between  $A$  and the  $k$  other individuals. The externalities are formally defined by the presence of one (or more) of the  $k$  individuals' levels of utility and/or consumption of any or all commodities in  $A$ 's utility function. Thus, letting a superscript attached to any variable denote possession, the utility functions of the two types of individuals can be written:

$$(1) \quad U^j = U^j(X_1^j, \dots, X_n^j) \quad j = 1, \dots, k$$

$$(2) \quad U^A = U^A(X_1^A, \dots, X_n^A; X_1^1, \dots, X_n^1; \dots; X_1^k, \dots, X_n^k; U^1, \dots, U^k)$$

The existence of the external effects is indicated by the inclusion of the quantities consumed of the various commodities by the  $k$

<sup>2</sup> For a thorough discussion of those conditions which must prevail for one-way transfers to take place, see Daly and Gietz.

<sup>3</sup> Consideration of cases where some goods and/or the utility levels of some individuals exerted external diseconomies would not alter the formal conclusions of this paper but would complicate the analysis and thus is not included.

individuals and by the separate inclusion of their utilities in equation (2).<sup>4</sup> The relevant external effects may appear in two different ways in  $A$ 's utility function. First,  $A$  may derive utility from another individual's consumption of specific commodities. For example, the external effect of housing, food, and education may be very substantial to  $A$ , while that of liquor and tobacco may be very close to zero (or, quite conceivably, negative). The effect of this kind of externality is accounted for by the inclusion of the  $X_1^1, \dots, X_n^1; \dots; X_1^k, \dots, X_n^k$  in  $A$ 's utility function. It is possible however that  $A$  also derives utility strictly from another individual's level of satisfaction regardless of how this is achieved. This is accounted for by the inclusion of  $U^j$ . If  $\partial U^A / \partial U^j$  is equal to zero and any or all of the  $\partial U^A / \partial X^j$  are positive, the external effect is designated as a *goods externality*. If the opposite is true, i.e.,  $\partial U^A / \partial U^j > 0$  and  $\partial U^A / \partial X_i^j = 0$ , the effect is labeled as a *utility externality*. If both partial derivatives are positive, both goods and utility externalities are present and if both are zero,  $A$ 's utility function is strictly private.<sup>5</sup>

### B. Externalities, Transfers, and Pareto Optimality

Given the existence of external effects of the types specified above, increments to the consumption of any of the individuals who exert external effects upon  $A$  may increase  $A$ 's utility. As has been noted elsewhere, should such an increase exceed the cost to  $A$  (in terms of utility lost in foregoing own consumption) he would then wish to make trans-

fers to some or all of the  $j$  individuals. (See Boulding (1962), Harold Hochman and James Rodgers, Mark Pauly.) Moreover, should such transfers have substantial "free-rider" effects (i.e., should recipients' welfare exert multi-party external effects) the transfers may be collectively organized. (See Knut Wicksell, Robert Goldfarb, Richard Zeckhauser.)

The purpose of this paper, however, is not to discern those conditions under which transfers take place but rather to consider the role of trade in achieving Pareto optimality when relevant consumption externalities exist. The importance of this issue resides in the fact that it is precisely the presumption that trade will lead to Pareto optimal states that lies behind the assertion that purchasing power transfers possess welfare superiority.

To derive the conditions necessary for Pareto optimality in exchange between person  $A$  and the remaining  $k$  members of society, the following Lagrangian function is formed:

$$(3) \quad L = U^A + \sum_{j=1}^k \alpha_j (U^j - \bar{C}^j) + \sum_{i=1}^n \beta_i \left( X_i^A + \sum_{j=1}^k X_i^j - \bar{X}_i \right)$$

Maximizing this function is equivalent to maximizing  $A$ 's utility subject to the constraint that the utilities of each of the other  $k$  persons remain constant ( $U^j = \bar{C}^j$ ) and the constraint that consumption of  $X_i$  equals the fixed supplies of the  $n$  goods. Rearrangement of the first-order conditions of (3) yields the following marginal conditions necessary for Pareto optimality between  $A$  and the other members of society:

$$(4) \quad \frac{\partial U^A}{\partial X_i^A} - \frac{\partial U^A}{\partial X_i^j} - \frac{\partial U^A}{\partial U^j} \frac{\partial U^j}{\partial X_i^j} = \frac{\partial U^j}{\partial X_i^j} \\ = \lambda_{i,h}^{A,j} \quad \begin{matrix} j = 1, 2, \dots, k \\ h = 1 \\ i = 2, 3, \dots, n \end{matrix}$$

<sup>4</sup> Since the utility functions of all the individuals except  $A$  are strictly private, reciprocal externalities and public goods externalities (e.g., cases where improvements in a recipient's welfare benefitted all other individuals) are not considered. The basic model could be extended fairly easily to take account of these features and such an extension would not alter our conclusions. Limitations of space preclude such a modification.

<sup>5</sup> Obviously, the only way the  $j$ th individual's utility can be changed in our model is by changing his consumption of some or all goods. Thus, it would be possible to include both the utility and goods externality in a single term. However, because of the vital importance of the distinction between the two types of externality to our conclusion, the effects are written in separate terms.

where the left-hand side represents  $A$ 's marginal condition in relation to the  $j$ th individual, the expression to the right of the first equality sign is the  $MRS$  of the  $j$ th individual and  $\lambda_{i,h}^{A,j}$  represents the price ratio which equates the two. The  $MRS$  of the  $j$ th individual is private, thus reflecting the nature of his preference function as shown in expression (1) above. It should also be noted that if only goods externalities exist, the third term in both the numerator and denominator of the left-hand side would vanish. If only utility externalities exist, the second terms in numerator and denominator would vanish. If neither type of externality is present both terms would vanish thus yielding the traditional necessary conditions for Pareto optimality in exchange.

Given external effects of the types specified above, will trade lead to results which are Pareto optimal? The answer to this question forms the central proposition of this paper. That answer is conditional and depends upon the form taken by the externalities. If there exist goods externalities (with or without utility externalities), competitive trade will not, in general, lead to Pareto optimality. If, on the other hand, only utility externalities are present (or if no externalities exist) trade is consistent with the achievement of Pareto optimality. The crucial issue, then, centers on whether external effects are generated by the consumption of particular commodities, i.e., the existence of goods externalities.

A fundamental factor underlying this conclusion is the nature of competitive trade. A prerequisite for such trade is the existence of large numbers of individuals engaging in impersonal exchanges. Such trade also implies the establishment of a single, market clearing price. Further, in a world of private (orthodox) preference functions individuals equate their marginal rates of substitutions to the resultant price ratio and to each other's  $MRS$  thus establishing the Pareto conditions.

In contrast, trade will not bring about the Pareto results if goods externalities exist and differ in intensity among alternative pairs of individuals (traders), e.g., the consumption of a poor man affects  $A$  more than the consumption of a rich man. Under such circumstances, the establishment of the Pareto

conditions would require that  $A$  trade at different price ratios with different individuals, the exact price ratio depending on the strength of the external effects generated. For example,  $A$  would wish to trade a good at a lower price to an individual whose consumption of that good extended strong external benefits to him than to one whose consumption did not, precisely in order to induce the former to consume more of that commodity and thus to extend external benefits to him. In terms of equation (4) this means that the value of  $\lambda_{i,h}^{A,j}$  would depend upon which of the  $k$  individuals  $A$  was being compared (trading) with. Thus, a necessary condition for Pareto optimality is price discrimination, a phenomenon inconsistent with competitive exchange. Conversely, the establishment of a single market price ratio is, under the existence of the assumed goods externalities, incapable of equating all individuals'  $MRS$  and hence of achieving optimality.<sup>6</sup> This problem does not arise if only utility externalities exist and the reason is straightforward: since  $A$  is, under such circumstances, not influenced by others' consumption mix but only by the utility they achieve, he need not concern himself by "bribing" individuals with price differentials.

The conclusions reached here—that the achievement of Pareto optimality through trade is consistent with the existence of utility externalities but not with goods externalities—can perhaps be seen more clearly if approached in a different way. Competitive trade is by its very nature a depersonalized process. For this reason it is reasonable to assume that such a process ignores the existence of interdependence between individuals. If this is so, trade will equate traders' marginal rates of substitution, or, in the context of our model:

$$(5) \quad \frac{\frac{\partial U^A}{\partial X_i^A}}{\frac{\partial U^A}{\partial X_h^A}} = \frac{\frac{\partial U^j}{\partial X_i^j}}{\frac{\partial U^j}{\partial X_h^j}} \quad j = 1, 2, \dots, k$$

<sup>6</sup> Indeed, Pareto optimality would, under these circumstances, require a system of taxes and/or subsidies which vary over both commodities and individuals (see J. De V. Graaff).

In this context the vital issue is now to determine under what conditions, if any, the establishment of condition (5) implies the satisfaction of condition (4). To do this, we rearrange equation (4)—the marginal conditions for Pareto optimality—to obtain:

$$(6) \quad \frac{\partial U^A}{\partial X_i^A} \frac{\partial U^j}{\partial X_h^j} - \frac{\partial U^A}{\partial X_i^j} \frac{\partial U^j}{\partial X_h^A} - \frac{\partial U^A}{\partial U^j} \frac{\partial U^j}{\partial X_i^j} \frac{\partial U^j}{\partial X_h^j} \\ = \frac{\partial U^A}{\partial X_h^A} \frac{\partial U^j}{\partial X_i^j} - \frac{\partial U^A}{\partial X_h^j} \frac{\partial U^j}{\partial X_i^A} - \frac{\partial U^A}{\partial U^j} \frac{\partial U^j}{\partial X_h^j} \frac{\partial U^j}{\partial X_i^j}$$

Subtracting the common terms from both sides of equation (6) yields:

$$(7) \quad \frac{\partial U^A}{\partial X_i^A} \frac{\partial U^j}{\partial X_h^j} - \frac{\partial U^A}{\partial X_i^j} \frac{\partial U^j}{\partial X_h^A} \\ = \frac{\partial U^A}{\partial X_h^A} \frac{\partial U^j}{\partial X_i^j} - \frac{\partial U^A}{\partial X_h^j} \frac{\partial U^j}{\partial X_i^A}$$

Now the issue boils down to this question: Under what conditions are expressions (5) and (7) equivalent? The answer is that they are equivalent if, and only if, equation (8) is satisfied.

$$(8) \quad \frac{\partial U^A}{\partial X_i^j} \frac{\partial U^j}{\partial X_h^j} = \frac{\partial U^A}{\partial X_h^j} \frac{\partial U^j}{\partial X_i^j}$$

Equation (8) will in turn be satisfied if:

$$(9) \quad \frac{\partial U^A}{\partial X_i^j} = \frac{\partial U^A}{\partial X_h^j} = 0$$

or, in other words, if no goods externalities exist. It will also be satisfied if:

$$(10) \quad \frac{\frac{\partial U^A}{\partial X_i^j}}{\frac{\partial U^A}{\partial X_h^j}} = \frac{\frac{\partial U^j}{\partial X_i^j}}{\frac{\partial U^j}{\partial X_h^j}}$$

that is, if  $A$ 's evaluation of the  $j$ th individual's consumption between  $X_i$  and  $X_j$  is the same as that person's own  $MRS$  between the two goods. In this unique and perhaps trivial case there exists no meaningful distinction between goods and utility externalities since  $j$  will make precisely those consumption choices  $A$  would wish him to make. The alge-

braic results thus confirm our initial statement—that trade is consistent with the achievement of Pareto optimality if no externalities exist, or the externalities are related only to levels of utility and not to consumption of specific commodities.

A crucial issue here is the nature of the trading process. For certain purposes, among them the achievement of utility maximization for the isolated individual, the depersonalized process called competitive trade is extremely efficient. Hence, in a private world in which individual preference functions are, in effect, isolated such trade is consistent with the collective achievement of Pareto optimality. As noted above, it is this world that economists who argue for purchasing power transfers apparently have in mind. On the other hand, as noted long ago, it is precisely this depersonalized characteristic of trade which makes it incapable of transmitting benevolent impulses among individuals (see Wicksell). In the context of our model, competitive trade *is not* an efficient method of maintaining a system of discriminatory prices. Yet, such a system *is* a necessary condition for Pareto optimality in a world of goods externalities.

### III. Theory and Policy

#### A. Money vs. Goods Revisited

Our analysis suggests that the form taken by consumption externalities is as vital an issue as their existence. Do people, individually or collectively, extend aid to others in hopes of improving the real welfare of the recipients or in order to alter their consumption patterns? It is instructive to note in this regard that the behavior of private, voluntary redistributive schemes presumably reflects the preferences of donors. We find that these institutions, while typically receiving income in the form of purchasing power, almost inevitably redistribute it in the form of particular commodities. In many cases, not surprisingly, the commodity redistributed is a perfectly illiquid service and the basis on which recipients are selected is not necessarily a low level of real income but a need for that particular service. For example, aid given to disease victims by private

charities is almost always in the form of free or subsidized treatment rather than money.<sup>7</sup> The nature of these institutions, therefore, would seem to suggest the predominance or at least existence of "goods" externalities in the utility functions of donors.

### *B. Transfers and Preferences*

Pareto optimality and the role of trade in achieving it, while of substantial interest to the economist, need not be particularly relevant to the formulation of actual economic policy. In terms of predicting and/or explaining observed behavior a more relevant issue is the kind of transfers desired by individuals (both donors and recipients) since it is these desires, expressed in ballot boxes and charitable contributions, that ultimately determine the nature and scope of welfare programs.<sup>8</sup> Our distinction between goods and utility externalities suggests that the ultimate motivation for nonrecipient support of public welfare schemes may lie precisely in the ability of those plans to influence consumption patterns. Thus, a majority of voters might approve of public housing for slum areas but be strongly opposed to the transfer of money to people living in those slums (see Olsen (1971)). For donors the crucial issue may be not the level of well being achieved by the recipient but rather *how* he achieves it.

Obviously, we would always expect recipients to prefer purchasing power transfers. In contrast, should goods externalities exist and should the tastes of donors and recipients differ concerning the recipients' consumption choices, the donors will prefer the redistribution of at least some sets of commodities to transfers of an equivalent sum of money. The two forms of transfer would be equivalent in

the donor's eyes only if: a) he and the recipients have identical tastes and income elasticities with respect to the recipients' consumption; and/or b) only utility externalities (or no externalities) exist. In all other cases the externally affected party will prefer the transfer of some sets of goods.

It should also be noted that, in addition to the tastes of donors and recipients, the liquidity or marketability of commodities will affect the choice of and form in which commodities are transferred in a world of goods externalities. Other things equal, donors subject to goods externalities will tend to favor the redistribution of illiquid rather than liquid commodities. This is because the transfer of highly liquid commodities approximates purchasing power transfers.

Most importantly, on a practical political level our analysis suggests that the efforts of many welfare reformers, however well-intentioned, may be misguided and lead to perverse consequences. Presumably much of the support for welfare reform of the type discussed in this paper comes from people who wish to improve the lot of welfare recipients. Yet, if it is the consumption patterns of recipients which motivate the political majority to support welfare assistance, such reformers may be naive in seeking changes in the current system. If only cash transfers were permitted and donors therefore precluded from influencing the consumption mix of recipients, it might well be that the real value of those transfers and the welfare of recipients would diminish. In this case everyone, donors as well as recipients, might suffer.

Given the democratic decision-making process it is not meaningful to discuss ways of improving recipients' welfare without considering those political and economic forces which determine the size of welfare transfers. Indeed, our paper suggests that these two issues—the size of welfare transfers and the way in which those transfers are allocated by recipients—may be vitally related since the latter may importantly influence the former. Economists and welfare reform advocates who consider these issues separately

<sup>7</sup> Even if one accepts the specialization of charitable institutions as resulting from "technical" factors (e.g., economies of scale, minimization of information costs), it does not follow that redistribution should take the form of specific goods and/or services. Those truly interested in the personal welfare of disease victims would still wish to give them money rather than drugs, operations, etc.

<sup>8</sup> The argument here is along the lines suggested by Buchanan.

may be making serious analytical and strategic errors.<sup>9</sup> Stating this proposition in its simplest form, welfare recipients' real alternatives may not be between a sum of money and a set of goods of equal market value but rather between a set of goods of one market value and a much smaller sum of cash. If this is so, neither compassion, political expediency nor Paretian welfare economics suggests an a priori preference for cash transfers.

#### IV. Conclusion

As we noted at the beginning of this paper, there are a number of reasons why economists might favor redistributions of purchasing power rather than specific and illiquid commodities. What we have shown here is that one of those reasons—efficiency in terms of achieving Pareto optimality—is not valid under certain highly plausible conditions regarding the existence and form of consumption externalities. In addition, we have argued that existing redistributive schemes, emphasizing as they do the transfer of goods rather than purchasing power, may well be natural outcomes of the democratic political process. After all, the majority frequently imposes its tastes on the minority—censorship and laws prohibiting the consumption of certain commodities, for example.<sup>10</sup> Is it surprising that this same phenomenon be reflected in the form taken by societal income redistribution?

Needless to say, our conclusion does not suggest that economists (or anyone else for

that matter) should argue in favor of redistribution in the form of goods. Even ignoring the second best problems of the real world, it is highly unlikely that any system of direct grants can be tailored to approximate the Pareto conditions. In addition, our paper has implicitly assumed that transactions are costless. In a world (such as ours) in which donors contribute money and recipients receive goods, the government (or other redistributive agency) must convert the money into the chosen goods and/or enforce the consumption patterns chosen by the political majority. The resulting costs of these transactions and of the bureaucratic structures required for them may prove greater than any benefits. Further, while we have shown that given the existence of goods externalities any potential donor will prefer the redistribution of *some* set of illiquid commodities to a money transfer, it does not follow that he will prefer *any* transfer of goods to a cash transfer of equal market value. This has important implications in terms of a democratic political decision-making process. For example, suppose that a person has tastes which differ from those of recipients but differ even more from those of the political majority. Under these circumstances a person who is subject to "goods" externalities might prefer cash transfers to the transfer of the particular set of commodities favored by the political majority and, for this reason, favor welfare reform which emphasizes purchasing power transfers.

Many economists may well argue that the policy implications of this paper are "paternalistic" and, therefore, ethically reprehensible to them. There is, of course, nothing illegitimate with such an opinion. But, assuming the existence and relevance of altruistic goods externalities, these economists must also then reject as ethically objectionable in this context the notion of Pareto optimality which is, after all, an ethical standard. In fact, it is an ethical standard which argues, among other things, that *everyone* "counts." Our results are directly implied by the fact that, in a world of externalities, *everyone* is a term inclusive of the

<sup>9</sup> These considerations may also be highly relevant to transfers between governmental entities. For example, a plan of "revenue sharing" in the form of unrestricted cash grants from the federal to the state governments may meet formidable political opposition because the political constituency of the former dislikes the prospective allocation of the grant that would be made by the latter.

<sup>10</sup> Indeed, the logic that purchasing power transfers are efficient would, when applied to the issue of taxation (an opposite type of one-way transfer), argue that lump sum income taxes should be used for purposes of sumptuary taxation. Thus, for example, the consumption of heroin would be discouraged by an income tax on heroin addicts.

donors as well as the recipients of welfare programs.

## REFERENCES

- K. Boulding, "Economics as a Moral Science," *Amer. Econ. Rev.*, Mar. 1969, 59, 1-14.
- , "Notes on a Theory of Philanthropy," in F. G. Dickinson, ed., *Philanthropy and Public Policy*, New York 1962.
- J. M. Buchanan, "What Kind of Redistribution Do We Want?," *Economica*, May 1968, 40, 185-90.
- and W. C. Stubblebine, "Externality," *Economica*, Nov. 1962, 34, 371-84.
- G. Daly and F. Giertz, "Benevolence, Malevolence, and Economic Theory," *Publ. Choice*, fall 1972, 13, forthcoming.
- J. De V. Graaff, *Theoretical Welfare Economics*, Cambridge 1967.
- R. S. Goldfarb, "Pareto Optimal Redistribution: Comment," *Amer. Econ. Rev.*, Dec. 1970, 60, 994-96.
- H. M. Hochman and J. D. Rodgers, "Pareto Optimal Redistribution," *Amer. Econ. Rev.*, Sept. 1969, 59, 542-57.
- E. O. Olsen, "A Normative Theory of Transfers," *Publ. Choice*, spring 1969, 6, 39-58.
- , "Subsidized Housing in a Competitive Market: Reply," *Amer. Econ. Rev.*, Mar. 1971, 61, 220-24.
- M. Pauly, "Efficiency in the Provision of Consumption Subsidies," *Kyklos*, Jan. 1970, 23, 33-57.
- G. Tullock, "The Social Rate of Discount and the Optimal Rate of Investment: Comment," *Quart. J. Econ.*, May 1964, 68, 332-36.
- W. Vickrey, "One Economist's View of Philanthropy," in F. G. Dickinson, ed., *Philanthropy and Public Policy*, New York 1962.
- K. Wicksell, "A New Principle of Just Taxation," in R. Musgrave and A. Peacock, eds., *Classics in the Theory of Public Finance*, New York 1958.
- R. Zeckhauser, "Optimal Mechanisms for Income Transfer," *Amer. Econ. Rev.*, June 1971, 61, 324-34.

# A "One Line" Proof of the Slutsky Equation

By PHILIP J. COOK\*

One focus of the usual classroom discussion of consumer theory is the demonstration that the individual consumer's reaction to a change in the market price of a commodity can be usefully broken down into vectors of substitution effects and income effects. The Slutsky equation relating the price effect to the substitution and income effects can be simply motivated by J. R. Hicks' graphical presentation, p. 31, but the usual proof (see Paul Samuelson) is very tedious and non-intuitive. If the instructor includes a discussion of the expenditure function in his curriculum, however, he has available a concise, intuitively appealing proof of the Slutsky equation.<sup>1</sup>

Suppose a consumer with income  $y$  faces a vector of commodity prices  $p$ . His Marshallian demand curve for commodity  $j$  is given by  $x_j = D^j(y, p)$ . The minimum expenditure necessary for the consumer to achieve any utility level  $u$  is given by his expenditure function,  $y = m(u, p)$  (here  $y$  is in units of the  $j$ th good). His Hicksian income-compensated demand for commodity  $j$  is represented  $x_j = h^j(u, p)$ ; if  $m$  is differentiable, we have the well-known result that

$$(1) \quad h^j(u, p) = \frac{\partial m(u, p)}{\partial p_j}$$

By the way the functions are defined, we

\* Graduate student, University of California, Berkeley.

<sup>1</sup> The expenditure function has been analyzed by L. McKenzie, S. Karlin, and D. McFadden and S. G. Winter, Jr.

have the identity

$$(2) \quad h^j(u, p) \equiv D^j(m[u, p], p)$$

Taking derivatives with respect to the price  $p_i$  of some commodity  $i$  yields, by the composite function rule:

$$(3) \quad \frac{\partial h^j(u, p)}{\partial p_i} = \frac{\partial D^j(y, p)}{\partial y} \cdot \frac{\partial m(u, p)}{\partial p_i} + \frac{\partial D^j(y, p)}{\partial p_i}$$

where  $y = m(u, p)$ . Using (1) and rearranging terms gives us the Slutsky equation:

$$(4) \quad \frac{\partial D^j(y, p)}{\partial p_i} = \frac{\partial h^j(u, p)}{\partial p_i} - x_i \frac{\partial D^j(y, p)}{\partial y}$$

noting again that  $y = m(u, p)$ .

## REFERENCES

- J. R. Hicks, *Value and Capital*, 2d ed., London 1946.
- S. Karlin, *Mathematical Methods and Theory in Games, Programming, and Economics*, vol. 1, Reading, Mass. 1959.
- D. McFadden and S. G. Winter, Jr., *Theory of Resource Allocation and Prices*, unpublished manuscript, Univ. California, Berkeley 1969.
- L. McKenzie, "Demand Theory Without a Utility Index," *Rev. Econ. Stud.*, June 1957, 24, 185-89.
- P. A. Samuelson, *Foundations of Economic Analysis*, New York 1967.

# A Geometric Treatment of Averch-Johnson's Behavior of the Firm Model: Comment

By ROBERT J. STONEBRAKER\*

In his March 1970 article in this *Review*, E. E. Zajac has provided us with a helpful and enlightening graphical exposition of the Averch-Johnson effect. Unfortunately he has made an error in claiming that maximizing the "reward to owners (stockholders)" leads us to an indeterminate solution.

The difficulty lies with his interpretation of this goal. Specifically he equates the rate of return to stockholders to the rate of return earned on equity capital. However, this is incorrect on two counts. The return to stockholders must be computed with respect to the market value of the equity rather than the book value and must include any capital gains (or losses) which may accrue. By incorporating these changes we can show that the result is determinate and is identical to that derived by Zajac for profit maximization. Of course, the indeterminacy still will hold for a firm maximizing the rate on equity capital, but this is a rather meaningless concept and of little value.

The problem faced by a regulated firm in our case is then to maximize

$$(1) \quad r_e = \frac{P_1 S_0 - P_0 S_0}{P_0 S_0} + \frac{E_1/S_1}{P_0 S_0} S_0$$

where  $P$  is the stock price,  $S$  is the number of shares outstanding,  $E$  is the earnings available to stockholders, and the subscripts on these variables refer to periods one and zero. Equation (1) can be interpreted as the percent capital gain plus the rate of return of earnings to each dollar of equity outstanding at the beginning of the period. The regulatory constraint as formulated by Zajac is

$$(2) \quad \pi \leq (f - i)K$$

\* Assistant professor of economics, Indiana University of Pennsylvania. The work was done while I was a graduate student at Princeton University.

where  $\pi$  is economic profit;  $K$  is total firm capital;  $f$  is the maximum allowable rate of return; and  $i$  is the average cost of capital. This can be rewritten as:

$$(3) \quad \frac{E_1 + i_a f_a K}{K} \leq f$$

or

$$(4) \quad E_1 \leq K(f - i_a f_a)$$

where  $i_a$  is the cost of debt capital;  $i_e$ , the cost of equity capital;  $f_a$ , the proportion of capital held as debt; and  $f_e$ , the proportion held as equity. The capital structure will be assumed to remain constant.

Now equilibrium in a competitive stock market requires that the ratio of earnings per share to price be equal for all firms with equal risk and equal expected growth (which we assume for convenience to be zero), and that this ratio is the cost of equity capital.<sup>1</sup> Thus for every period we have

$$(5) \quad i_e = \frac{E/S}{P}$$

Substituting into equation (1) we get

$$(6) \quad r_e = \frac{P_1 S_0 - E_0/i_e}{E_0/i_e} + \frac{E_1/S_1}{P_0} \\ = \frac{E_1 S_0/i_e S_1 - E_0/i_e}{E_0/i_e} + \frac{E_1/S_1}{E_0/i_e S_0} \\ (7) \quad = (1 + i_e) \frac{E_1/S_1}{E_0/S_0} - 1$$

Since everything but  $E_1$  and  $S_1$  is constant, we need only concern ourselves with the earnings per share. The firm wants to get  $E_1/S_1$  as high as is consistent with its constraint. If we can prove that the firm can

<sup>1</sup> See James C. Van Horne, ch. 6, for a discussion of these concepts.

raise  $E/S$  by adding more capital with any given rate of return, then we know that the firm will use the maximum amount of capital for each rate; a situation which Zajac has shown to be one of overcapitalization.

Now an increase of  $\Delta K$  requires  $f_e \Delta K$  dollars of new equity or  $f_e \Delta K/P$  shares. Furthermore, the firm can make  $f \Delta K$  in profits with this or  $\Delta K(f - i_d f_d)$  in earnings. If  $(E + \Delta E)/E$  is greater than  $(S + \Delta S)/S$  then we can conclude that the earnings per share increases as capital increases for a given rate of return. Thus we need

$$(8) \quad \frac{E + \Delta K(f - i_d f_d)}{E} > \frac{S + f_e \Delta K/P}{S}$$

$$\frac{\Delta K(f - i_d f_d)}{E} > \frac{f_e \Delta K}{PS}$$

$$f - i_d f_d > E f_e / PS = i_e f_e$$

$$(9) \quad \frac{f - i_d f_d}{f_e} > i_e = \frac{i - i_d f_d}{f_e}$$

This is true, of course, whenever  $f > i$  as is

postulated by Zajac. The left-hand side of (9) is simply the marginal return to stockholders and the right side the cost of equity capital. We are left with the reasonable conclusion that when the marginal revenue is greater than cost, the firm has an incentive to expand its capital.

Whereas Zajac concluded the firm is indifferent among all input combinations yielding a return of  $f$ , it seems apparent that, as long as  $f$  is greater than  $i$ , the firm will maximize the amount of capital it uses in obtaining a rate of return of  $f$  percent.

# REFERENCES

- H. Averch and L. L. Johnson, "Behavior of the Firm Under Regulatory Constraints," *Amer. Econ. Rev.*, Dec. 1962, 52, 1053-69.
- J. C. Van Horne, *Financial Management and Policy*, Englewood Cliffs 1968.
- E. E. Zajac, "A Geometric Treatment of Averch-Johnson's Behavior of the Firm Model," *Amer. Econ. Rev.*, Mar. 1970, 60, 117-25.

# A Geometric Treatment of Averch-Johnson's Behavior of the Firm Model: Reply

By E. E. ZAJAC\*

Robert Stonebraker argues that a model of the regulated firm should have management trying to maximize stockholders' total return: dividends (presumed to be earnings) plus capital gain. This leads one to the capital markets and to theories involving difficult-to-measure expectations and behavior of both investors and regulators. Regardless of theoretical assumptions about the capital markets and what regulators and managers *should* do, one observes that they pay great attention to rate of return, a quantity that is relatively easily determined and monitored. It would seem premature to dismiss rate of return as "a rather meaningless concept of little value."

If one accepts Stonebraker's model, it is clearly untenable to make his assumption that the firm is guaranteed earnings at an allowed rate  $f$ , greater than the cost of capital,  $i$ . To conclude that such a firm will have a tendency to expand because it will make  $\Delta K(f - i_d f_d)$  in equity earnings on  $\Delta K$  is analogous to concluding that the profit-maximizing firm will expand because it is guaranteed profits of  $\Delta \pi = (f - i)\Delta K$  on  $\Delta K$ . In either case the firm would expand indefinitely (indeed one would assume that the regulators would eventually set  $f = i$  to bring the firm's return in line with the opportunity

cost of capital). The point is that the constraint of  $f$  is merely an upper bound and not a guaranteed earnings rate. Earnings possibilities as well as the constraint must be considered to make resource allocation conclusions. To consider these possibilities in Stonebraker's model, one would write the firm's  $E/S$  maximand as

$$\frac{E}{S} = \frac{R - wL - i_d f_d(K + \Delta K)}{S_0 + (f_e \Delta K / P_1)}$$

where  $L$ ,  $w$  are labor and the wage rate,  $R = R(K + \Delta K, L)$  is the revenue function, and  $K$  and  $S_0$  are the firm's capital and number of shares of stock before expansion. If  $P_1 > f_e K / S_0$  and if the cost of equity capital,  $i_e$ , satisfies  $i_e \geq (E/S) / P_1$ , it can be shown that maximization of  $E/S$  subject to the regulatory constraint (Stonebraker's equation (3)) yields over-intensive capital substitution. For  $i_e < (E/S) / P_1$  either over- or under-intensive capital substitution can result. If one assumes an equilibrium wherein investor's expectations of  $E/S$  and the firm's maximization of  $E/S$  converge to values of  $E/S$  and  $P_1$  that satisfy  $(E/S) / P_1 = i_e$ , then Stonebraker's model in fact gives a result equivalent to that of Averch-Johnson. However, without empirical studies it is not clear whether a profit maximand, an  $E/S$  maximand and perfect capital markets, or a rate of return maximand most accurately describes the regulated firm.

\* Bell Telephone Laboratories, Inc. The views are solely those of the author and are not necessarily opinions of the Bell Telephone Laboratories nor of the American Telephone and Telegraph Company.

# Security Pricing and Investment Criteria in Competitive Markets: Comment

By PREM KUMAR\*

Recently, Jan Mossin presented a security pricing model within the framework of a market equilibrium theory. The model is based on particular preference structures of investors, specified in terms of quadratic utility functions with final wealth as the argument of the functions. As an implication of his model for the firm's optimal investment policy, Mossin demonstrates how Proposition III put forth by Franco Modigliani and Merton Miller (1958) can be validated. In addition, the analysis is extended to suggest investment criteria for investments with completely arbitrary yield characteristics.

The purpose of this comment is twofold. First, to show that Mossin's proof of the validity of M-M's Proposition III is questionable, given his assumptions. Second, an attempt is made to show how a troublesome assumption of Mossin's analysis could possibly be eliminated.

For the sake of exposition, Mossin's security pricing model as shown on page 752, equation (6), is stated below with all relevant definitions:

$$(1) \quad v_j = \frac{1}{r} \left[ \mu_j - \frac{\sum_k \sigma_{jk}}{\sum_i \frac{1}{2c_i} - \sum_k \mu_k} \right]$$

where

$r$  = the certain rate of return (see p. 752, fn. 4)

$X_j$  = gross yield (with  $E(X_j) = \mu_j$ );

$\sigma_{jk} = E[(X_j - \mu_j)(X_k - \mu_k)]$ ;

$d_j$  = debt of company  $j$ ;

$p_j$  = market value of shares of company  $j$ ;

$v_j = p_j + d_j$  = market value of company  $j$ ;

\* Assistant professor of finance, University of Massachusetts. I am grateful to Pao Cheng and James Ludtke for their helpful comments and suggestions.

$c_i$  = the risk aversion coefficient in  $i$ th investor's utility function<sup>1</sup> ( $c_i > 0$ ,  $Y_i \leq C_i/2$ )

The sum of the covariances of  $j$ th company yield with the yields of all companies participating in the market is represented by

$$b_j = \sum_k \sigma_{jk}$$

This, in the case of quadratic utility functions, is the relevant risk measure associated with company  $j$ 's activities; it can be interpreted as the contribution of company  $j$  to the market's total risk,  $\sum_j b_j$ . Mossin notes: "It is not very surprising to learn that a company's risk cannot be measured by its own variance ( $\sigma_{jj}$ ) alone, but also depends on its correlation with other firms" (p. 752).

Further, the expression in equation (1), which serves as a weighting factor for the company risk  $b_j$ , is<sup>2</sup>

$$(2) \quad R = \frac{1}{\sum_i \frac{1}{2c_i} - \sum_k \mu_k} \equiv \frac{1}{C/2 - \mu}$$

$R$  is the same for all companies and can be regarded as market risk aversion. Put another way,  $R$  is a sort of market discount

<sup>1</sup> From the quadratic utility functions (see Mossin, p. 752), since marginal utility must be positive everywhere, it follows that

$$U'_i(Y_i) = 1 - 2c_i Y_i \geq 0$$

where  $Y_i$  is the final wealth for investor  $i$ . Further, diminishing marginal utility implies  $-2c_i < 0$ , i.e.,  $c_i > 0$ . We call  $1/c_i (= C_i)$  the "risk tolerance" of investor  $i$ .

<sup>2</sup> For simplicity we define:

$$\sum_k \mu_k = \mu$$

and

$$1/2 \sum_i 1/c_i = 1/2 \sum_i C_i = C/2$$

factor for determining the risk premium,  $b_j R$ , which is deducted from the expected yield,  $\mu_j$ , to arrive at the certainty equivalent yield of company  $j$ .<sup>3</sup> From equation (1) it is clear that this certainty equivalent yield is discounted at the riskless rate,  $r$ , in order to determine the present market value of company  $j$ .

When considering the M-M investment theory, Mossin correctly specifies that

$$X'_j = (1 + \lambda)X_j,$$

and, therefore,

$$\mu'_j = (1 + \lambda)\mu_j,$$

for the firm  $j$  to stay within the same risk class.<sup>4</sup> But the statement regarding the sum of *all* covariances, when yield  $X_j$  increases by 100  $\lambda$  percent in every state of the world,

$$b'_j = (1 + \lambda)b_j$$

is not correct. The reason is:

$$\sigma'_{jj} \neq (1 + \lambda)\sigma_{jj}$$

Given the straightforward definition of variance, it follows that

$$\sigma'_{jj} = (1 + \lambda)^2 \sigma_{jj} = (1 + \lambda)\sigma_{jj} + \lambda(1 + \lambda)\sigma_{jj}$$

At this point, it should be noted that

$$\lambda(1 + \lambda)\sigma_{jj} \neq 0,$$

because it is the product of positive terms:  $\lambda > 0$  and  $\sigma_{jj} > 0$ .<sup>5</sup> Therefore, Mossin's expression for  $b'_j$  should be rewritten as follows:

$$(3) \quad b'_j = (1 + \lambda)b_j + \lambda(1 + \lambda)\sigma_{jj}$$

<sup>3</sup> See John Pratt, p. 125. In equilibrium, market risk aversion equals the "market price of risk" (see John Lintner).

<sup>4</sup> See Mossin, pp. 753-54. Also, see Modigliani-Miller (1958, p. 266). According to the M-M definition of a risk class—brought out more lucidly and precisely in 1963—the random variables  $X'_j/\mu'_j$  and  $X_j/\mu_j$  must be identically distributed. This condition is satisfied.

<sup>5</sup> If it is argued that  $\lambda$  is negligibly small, then the purpose of investment analysis is defeated, because this is tantamount to saying that either investment is negligible, or that investment is ineffective in changing the yield significantly. Further, if it is argued that  $\sigma_{jj}$  is negligible, then  $\sigma_{jk}$ 's (for all  $k$ ) would also be negligible, which would seem to suggest that firm  $j$  be branded as a non-risky firm.

Equation (3) leads to some interesting implications in reference to Mossin's general equilibrium model and the M-M investment theory. In the first place, when investment analysis satisfying the risk class assumption of the M-M theory is considered, the definition and measurement of risk according to quadratic utility preference structures shifts the firm into a different risk class in the sense that  $b'_j$  does not increase by 100  $\lambda$  percent but increases by more than that. This nonproportional increase in the company's risk measure suggests the disturbing conclusion that M-M's definition of risk (class) and the risk measure according to quadratic utility functions are logically inconsistent.

M-M's position on this point may be assessed from their statement on the validity of arbitrage proof, which is heavily dependent on their definition of a risk class:

The trouble stems mainly . . . on restating our proof in terms of the variances of returns to the arbitrating shareholder. Why anyone would want to introduce the variance of returns into the proof is particularly hard for us to understand since the essential motivation underlying our original arbitrage proof, after all, was precisely to avoid having to establish tradeoffs between the moments of the probability distribution of returns. [1969, p. 592]

From the quotation, then, it seems clear that M-M are referring to a much more general and broader definition of risk than the one captured by the covariance of returns.

Furthermore, after assuming that the number of firms is so large that the change in the risk aversion,  $R$ , is negligible, Mossin notes:

Under this assumption it is easy to see from the first-order conditions (4) that the proportionality relation holds for arbitrary utility functions: if a change in one company's yield does not appreciably effect the investor's marginal utility, then equations (4) are satisfied only if  $v_j$  and  $X_j$  are changed in the same proportion. [p. 754]

The statement, in my judgment, implies that with general investor preferences

$$v'_j = (1 + \lambda)v_j$$

when  $X_j$  increases by 100  $\lambda$  percent in all states of the world. But if Mossin's equations (4) are satisfied,<sup>6</sup> say for quadratic utility preference structures of investors, then it also follows from (1) and (3) that

$$\begin{aligned} v'_j &= \frac{1}{r} [\mu'_j - b'_j R] \\ &= (1 + \lambda)v_j - [(1/r)\lambda(1 + \lambda)\sigma_{jj}R] \end{aligned}$$

Obviously, there is some ground to question the validity of Mossin's generalization.

What is more, the change in one company's yield would necessarily change final wealth, thus affecting investors' marginal utilities—except in the trivial case of linear utility functions. This simple result negates the assumption invoked in Mossin's generalization. Hence, it is very difficult to justify the proportionality relation—even as an approximation.

Insofar as Mossin's proportionality relation is suspect, the validation of M-M's Proposition III is also suspect. Using, for example, debt financed investment at a cost of  $I$  which will increase yield by 100  $\lambda$  percent in all possible states, the new level of debt becomes

$$d'_j = d_j + I$$

Then the market value of new equity, even if  $R$  remains constant (following the increase in  $\mu_j$ ), is

$$\begin{aligned} p'_j &= v'_j - d'_j \\ &= p_j + \lambda v_j - [(1/r)\lambda(1 + \lambda)\sigma_{jj}R] - I \end{aligned}$$

It follows that the investment will be acceptable only if

$$\lambda v_j \geq [(1/r)\lambda(1 + \lambda)\sigma_{jj}R] + I$$

Now the following observation can be

<sup>6</sup> The first-order conditions (4) by Mossin are reproduced below:

$$\frac{\partial E[U_i(Y_i)]}{\partial z_{ij}} = E[U'_i(Y_i)(X_j - rv_j)] = 0 \quad (\text{all } j)$$

where  $Y_i = ru_i + \sum_j z_{ij}[X_j - rv_j]$ ;

$w_i$  = initial wealth for investor  $i$ ;

and  $z_{ij}$  = fraction owned of company  $j$  by investor  $i$

made effortlessly: the magnitude of  $[(1/r)\lambda(1 + \lambda)\sigma_{jj}R]$ , no matter how small, is instrumental in destroying the equality sign between  $\lambda v_j$  and  $I$ , and forces the strict inequality. Therefore, the investment will be acceptable if

$$\lambda v_j > I$$

or, if

$$\frac{\lambda \mu_j}{I} > \frac{\mu_j}{v_j} = \text{market capitalization rate for company } j$$

which refutes the M-M Proposition III, regardless of the method of financing the investment.

Next we address ourselves to the task of eliminating from Mossin's analytical framework a crucial assumption which plagues his investment analysis throughout. The assumption is that changes in  $R$ , because of changes in yields, are negligible.<sup>7</sup> It is desirable to dispense with this assumption since it bears on Mossin's investment criterion for a more general case of investments with completely arbitrary yield characteristics. More important,  $R$  must remain unchanged whether one firm undertakes investment or whether a number of firms undertake investments simultaneously. The following derivation is presented to accomplish this goal.

From equation (2),  $dR=0$ , only if

$$(4) \quad \frac{dC}{d\mu} = - \frac{\partial R / \partial \mu}{\partial R / \partial C} = 2$$

The economic interpretation of equation (4)

<sup>7</sup> Frankly speaking, "negligible" is an imprecise term not suited to precise economic analysis. At any rate, Webster's Dictionary defines negligible as something "that can be neglected or disregarded because small." But it is not made clear how small something has to be before discarding it, or in which precise situation something may be ignored. In the present context, there is certainly a dearth of support in Mossin's treatment that the deviation in  $R$  is minor and of the second order. There seems to be a big gap between the foregoing statement and the assumption that the number of firms is so large that the change in the risk aversion is negligible. And the assumption that the companies regard themselves as pricetakers in the security markets is of no help either. For instance, if a large number of firms undertook investments simultaneously, would the change in  $R$  still be negligible?

is that for any positive change in the aggregate market yield,  $\mu$ , the market risk tolerance,  $C$ , should also increase to cause small changes in the market risk aversion,  $R$ . Stated differently, if firms are undertaking investments which would increase the expected yields, then to make changes in  $R$  small or even zero, simply seek a wider distribution of corporate securities among investors. It may be clear that if the number of investors increase, the market risk tolerance,  $C$ , which is a function of the number of investors, also increases, thus offsetting the increase in  $\mu$ . Condition (4) specifies that, in order to obtain an unchanging equilibrium level of  $R$ , enough new investors ought to be attracted to offset exactly twice the increase in  $\mu$ .<sup>8</sup>

Condition (4) seems to be plausible in the author's opinion. In support of this it may be argued that inasmuch as real investment

<sup>8</sup> The effect of an increase in the number of investors on security valuation, through the effect of a change in the magnitude of market risk tolerance on the market price of risk, has been pointed out by writers, namely, Mossin, p. 753, and Lintner, pp. 95-96. This observation renders support to the acceptability of condition (4).

Realistically, the validity of (4) is not unequivocal because of the dependence on a restrictive preference structure. The point, though, is the indication of a line of reasoning that may be extended to encompass more general preference structures. In fact, whether or not (4)—or a similar more general condition—holds, is an interesting empirical question about the behavior of investors in a dynamic economy.

by a firm is triggered because of an expectation of a growing demand for the firm's products, or simply because of growth in population, it seems quite reasonable to assume the growth in investors, too. That is, in a growing economy, investors would participate in increasing numbers to equilibrate the capital market. Hence, we may be on a firmer ground to assume that condition (4) is satisfied, which maintains the expected risk aversion of the capital market as a whole at an unchanging level.

#### REFERENCES

- J. Lintner, "The Market Price of Risk, Size of Market and Investor's Risk Aversion," *Rev. Econ. Statist.*, Feb. 1970, 52, 87-99.
- F. Modigliani and M. H. Miller, "The Cost of Capital, Corporation Finance and the Theory of Investment," *Amer. Econ. Rev.*, June 1958, 48, 261-97.
- and ———, "Corporate Income Taxes and the Cost of Capital: A Correction," *Amer. Econ. Rev.*, June 1963, 53, 433-43.
- and ———, "Reply to Heins and Sprenkle," *Amer. Econ. Rev.*, Sept. 1969, 59, 592-95.
- J. Mossin, "Security Pricing and Investment Criteria in Competitive Markets," *Amer. Econ. Rev.*, Dec. 1969, 59, 749-56.
- J. W. Pratt, "Risk Aversion in the Small and in the Large," *Econometrica*, Jan.-Apr. 1964, 32, 122-36.

# Security Pricing and Investment Criteria in Competitive Markets: Reply

By JAN MOSSIN\*

I am grateful for the opportunity to correct and clarify my discussion of the Modigliani-Miller investment theory that has been provided by Prem Kumar's lucid comment. It is clear that my original discussion contained an error; what now seems most interesting is the interpretation and implications of the correct version of the analysis.

The pitfall that I fell into is the following: it is obviously true that if two firms,  $j$  and  $k$ , have proportional yields, i.e.,  $X_k = \alpha X_j$ , then company values stand in the same proportion to each other:  $v_k = \alpha v_j$ . As many others must have done before, I then jumped to the conclusion that if a company undertakes an investment which increases its yield by a certain percentage in all states of the world, i.e., changes  $X_j$  to  $X'_j = (1 + \lambda)X_j$ , then its value will increase in the same proportion. This, however, is not generally true. Such a proportionality relation can hold only as an approximation. Investor  $i$ 's  $j$ th demand function now changes from

$$\varepsilon[U'_i(Y_i)(X_j - rv_j)] = 0$$

to

$$\varepsilon[U'_i(Y'_i)((1 + \lambda)X_j - rv'_j)] = 0$$

where  $Y'_i$  and  $v'_j$  are the post-investment values of  $Y_i$  and  $v_j$ . If now  $Y'_i = Y_i$  we would obviously have  $v'_j = (1 + \lambda)v_j$ . However,  $Y_i$  is not entirely unaffected by changes in yields, since  $Y_i$  depends directly on aggregate yield  $\Sigma X_k$ , and this will necessarily change when the yield of any one company changes. Therefore, one way of "saving" the M-M investment theory is to treat this change as a minor and second-order effect by assuming that the number of firms is so large as to make the changes in marginal utilities negli-

gible. Such an assumption simply serves to define the approximation that is admitted by the proportionality assumption.

The considerations so far made no assumptions about investor preferences. I then proceeded to examine the proportionality relation when investors are assumed to choose mean-variance efficient portfolios, in which case equilibrium company values had been shown to be given by

$$v_j = \frac{1}{r} (\mu_j - Rb_j)$$

Here  $R$  is the market risk aversion factor and can be interpreted as the "price of risk" in the sense of being a measure of the marginal rate of substitution between expected yield and risk.

When, as a result of undertaking an investment, company yield changes from  $X_j$  to  $X'_j = (1 + \lambda)X_j$ , it is easy to see that the new expected yield is

$$\mu'_j = (1 + \lambda)\mu_j,$$

while for the covariance terms we have, as pointed out by Kumar,

$$\sigma'_{jk} = \begin{cases} (1 + \lambda)\sigma_{jk} & \text{for } k \neq j \\ (1 + \lambda)^2\sigma_{jj} & \text{for } k = j \end{cases}$$

The important implication of this is that the company's risk measure does *not* change in the same proportion as yield, but rather to

$$b'_j = (1 + \lambda)b_j + \lambda(1 + \lambda)\sigma_{jj}$$

Thus we always have  $b'_j \geq (1 + \lambda)b_j$ . It is now clear that because of the nonproportional change in  $b_j$ , the proportionality relation for company value will also be destroyed: instead of getting  $v'_j = (1 + \lambda)v_j$ , we obtain

$$v'_j = (1 + \lambda)v_j - \frac{1}{r} R' \lambda (1 + \lambda) \sigma_{jj}$$

\* Norwegian School of Economics and Business Administration.

where  $R'$  is the postinvestment value of  $R$ .

Earlier we were able to identify the proportionality assumption in terms of an assumption of marginal utilities being unaffected by the investment in question. We now see that the market valuation formula serves to define the approximation admitted by the proportionality assumption in a quite different way: the difference between the true change in company value and the change implied by the proportionality assumption comes about because of the nonproportional change in the company's risk measure  $b_j$ . Thus, calculations based on the proportionality assumption always overestimate the increase in company value. Another way of expressing this is to say that the firm faces a downward sloping demand schedule for its securities. It is important to keep in mind that a company with uncertain yield is in a way similar to the traditional monopolistic competitor simply because different shares are not perfect substitutes for each other. The difference is that this cannot be said to represent any lack of perfect competition in the capital market or to imply nonoptimality of the market's risk allocation.

The remaining problem is what assumptions we make about the effect of the change in the price of risk,  $R$ , caused by the investment. What basically seems to be the problem is what we prefer to mean when we characterize the firm as a "price taker." The proportionality assumption of the M-M investment theory can be considered as an assumption of a rather crude sort of price taking behavior for the firm, while a more sophisticated firm would be one which recognized the nonproportional change in  $b_j$  but still acted as a price taker with respect to  $R$ .

In many ways, the latter conception of the firm seems the more attractive.

For one thing, the ability to separate changes in  $b_j$  corresponds to an ability to distinguish between, on the one hand, "general market forces" represented by  $R$  and on which the firm may quite realistically believe it has no influence and, on the other hand, the unique characteristics of its own shares. Price taking with respect to  $R$  is from this point of view a very natural analogue in the case of uncertainty of the price taking behavior assumed in the traditional theory of commodity markets.

Another reason why the proportionality assumption is undesirable is that if an investment opportunity is characterized by constant or increasing stochastic returns to scale and the firm calculates on the basis of the proportionality assumption, then no finite equilibrium investment level exists. This is of course completely parallel to the inconsistency between nondecreasing returns to scale and perfect competition in traditional market theory. No such problems need arise when  $R$  is considered constant, however.

The last, but perhaps most important argument in favor of the assumption that firms act as price takers with respect to  $R$  is connected with the extension of the analysis to investments which are not of the nondiversifying M-M type. When a firm undertakes an investment with a yield pattern different from its current yield, it does not simply produce "more of the same," and in such a situation the proportionality assumption simply becomes meaningless. Furthermore, in the context of our market model, it seems that the only sensible way of specifying price taking behavior is with respect to the market risk aversion factor  $R$ .

# Macroeconomics of Unbalanced Growth: Comment

By MICHAEL KEREN\*

An unfortunate error of misinterpretation mars some of the conclusions of William Baumol's illuminating paper on the "Macroeconomics of Unbalanced Growth."

The paper develops a model of a two-sector economy: a "progressive" sector where labor-saving technical change takes place, and a nonprogressive sector where there is no technical change. His production functions are<sup>1</sup>

$$(1) \quad Y_{1t} = aL_{1t}$$

$$(2) \quad Y_{2t} = bL_{2t}e^{rt}$$

where  $Y_{jt}$  and  $L_{jt}$  stand, respectively, for output and labor input at time  $t$ , and the subscript  $j=1, 2$  refers to the unprogressive and progressive sectors respectively. With labor the only input, (given perfect competition), the price ratios of the outputs always equals the labor cost ratio. Letting  $W_t$  represent the wage rate,

$$(3) \quad \frac{P_{1t}}{P_{2t}} = \frac{W_t L_{1t} / Y_{1t}}{W_t L_{2t} / Y_{2t}} = \frac{L_{1t}}{L_{2t}} \cdot \frac{Y_{2t}}{Y_{1t}}$$

If we suppose that the elasticity of demand for the outputs were unitary, the ratio of outlays on both outputs would remain constant:

$$(4) \quad \frac{P_{1t} Y_{1t}}{P_{2t} Y_{2t}} = \frac{W_t L_{1t}}{W_t L_{2t}} = \frac{L_{1t}}{L_{2t}} = A (\text{constant})$$

The output ratio will then be

$$(5) \quad Y_{1t} / Y_{2t} = (aL_{1t}) / (bL_{2t}e^{rt}) = (aA / be^{rt})$$

which declines toward zero as  $t$  grows.

This result is interpreted in *Proposition 2*: "... There is a tendency for the outputs of the 'nonprogressive' sector ... to decline and perhaps, ultimately, to vanish" (Baumol, p. 418).

It is this *proposition* which is false. The

\* Lecturer in economics, Hebrew University, Jerusalem.

<sup>1</sup> The equations, and their numbering, are not identical with those of Baumol.

output of the nonprogressive sector 1, per capita of the working population, is *constant*, rather than declining. The output ratio of (5) falls because the denominator is rising, not because the numerator is falling. This becomes obvious by reference to equation (4). Assume that the total input of labor remains constant: the number of workers employed in each sector would also stay constant, and a constant input of labor produces, by equation (1), a constant output in the "nonprogressive" sector.

Several of Baumol's conclusions depend on the false Proposition 2. Thus, if the provision of higher education belongs to sector 1, and if the wage level is geared to the rising productivity in sector 2, it is true that the cost of higher education will rise (see p. 421). It does not follow that the relative outlays on higher education should rise, unless relatively more or better education is being provided. The same is true for municipal services. Let  $R$  be municipal revenues and  $L_1$  labor inputs into municipal services. Then the proportion of municipal deficits to municipal spending is

$$(6) \quad \begin{aligned} \frac{W_t L_{1t} - R_t}{W_t L_{1t}} &= 1 - \frac{L_t}{L_{1t}} \cdot \frac{R_t}{W_t L_t} \\ &= 1 - \frac{A + 1}{A} \cdot \frac{R_t}{W_t L_t} \end{aligned}$$

This proportion will rise only if the share of municipal revenues in total income ( $R/WL$ ) falls over time. In other words, only if the income elasticity of municipal taxes is below unity will the financial problems of cities grow, provided the level of services remains unchanged. This could happen if the municipal tax base does not grow as fast as labor income.

## REFERENCE

- W. J. Baumol, "Macroeconomics of Unbalanced Growth," *Amer. Econ. Rev.*, June 1967, 57, 415-26.

# Macroeconomics of Unbalanced Growth: Reply

By WILLIAM J. BAUMOL\*

There is no doubt that Michael Keren is right and that the point is important.<sup>1</sup> In the initial discussion of my model, I simply misinterpreted the rising relative cost of the urban public services to mean that it will become harder for society to provide them. As Keren shows, the rising productivity elsewhere in the economy that is the source of the increasing opportunity cost of the services, also automatically means that the community will be able more easily, if it wishes, to pay for these services, despite their rising cost.

The implications of Keren's point are worth spelling out. The basic argument of the original analysis still remains valid: the financial problem of the cities increases (in part) because the costs of the services rise more rapidly than the general price level. This may well lead to cumulative deteriora-

tion in the quality and quantity of the services, even though it is not forced on the economy by lack of resources.

Because of their rising relative costs, the community may in fact *want* less of some or even most of these services. But the danger is that the nature of the political process and the tax system will also force it to accept a decline in even those services which the public would in fact prefer to expand, an expansion which, as Keren shows, society will well be able to afford. The problem is to find means by which the rising resources of the community can be channelled into *those* public services at a rate that grows more rapidly than their costs.

## REFERENCES

- D. Bradford, "Balance on Unbalanced Growth," *Zeitschrift für Nationalökonomie*, 1969, 29, 291-304.
- I. K. Lynch and E. L. Redman, "Macroeconomics of Unbalanced Growth: Comment," *Amer. Econ. Rev.*, Sept. 1968, 58, 152-55.

\* Princeton University.

<sup>1</sup> My error has also been pointed out by several others, though sometimes not so clearly as in Keren's note. See, e.g., illuminating papers by David Bradford and by L. K. Lynch and E. L. Redman.

# Allais' Restatement of the Quantity Theory of Money: Note

By J. L. SCADDING\*

In a 1966 article in this *Review*, Maurice Allais presented a sophisticated and very successful method for estimating the demand for money. It departed from the usual investigations in (i) using the expected rate of change of outlays, which were assumed to be in fixed proportion to nominal output, rather than the expected rate of change of prices; and in (ii) using a time-variable distributed lag in the estimation of the expected rate of change of outlays.<sup>1</sup>

This note concentrates on analyzing that distributed lag and its use in specifying the demand for money. However, the results of the analysis suggest that the question of what is the correct argument in the demand function for money, expected rates of change of prices or outlays, is not independent of the specification of how they are estimated. As Phillip Cagan has noted (1969, p. 428), the crucial feature of Allais' distributed lag is that the weighting pattern "... rises and falls with velocity." The danger in this is that the expected rate of change of outlays computed from that distributed lag is used to estimate velocity.<sup>2</sup> The Allais procedure,

\* Assistant professor of economics, Stanford University. The research for this article was financed under National Science Foundation Grant GS-2530. I should like to express my indebtedness to Mordecai Kurz for his invaluable help and to Edward Shaw for his penetrating criticisms. Of course the usual waivers of responsibility apply, and the errors remain my property.

<sup>1</sup> Of course, in the case of hyperinflations the two will be the same. Allais is unique also in using a logistic curve formulation for the demand-for-money function. Allais notes (1969b, p. 443), however, that this is not crucial for his results. In any case, we can always assume that the same functional form is used no matter what the choice of arguments or specification of the estimation procedure.

<sup>2</sup> Michael Darby comments that velocity "... is being worked very hard: it serves as the demand function for money, the adjustment in the forgetfulness coefficient and the index of relative psychological time" p. 446. The important point here is that the latter two variables are used to estimate expected rates of change

therefore, may come down to regressing velocity on its past values. But if this is the case, whether rates of change of prices or of outlays is used is relatively unimportant; either "washes out" in the estimation process. Hence it would not be true as Allais claims that "... it is possible to choose ... between two different approaches ... only by confronting them with reality" (1969b, p. 444). The fact that extrapolations of time-series often give good predictions may be enough to explain the good results that Allais obtains.

The remainder of this note is taken up with showing how Allais' formulation of the distributed lag, and the definitions of the variables in it, lead to a method of predicting velocity which is essentially an extrapolation of velocity, and its derivatives, appropriately smoothed.

The notation used is that in the original Allais piece (1966); numbered references in parentheses are to equation numbers in that work. We denote the demand for nominal money balances per dollar of transactions as  $\phi_d$ . Transactions are assumed to be in fixed proportion to nominal output so that we identify  $\phi_d$  with the inverse of income velocity,  $V$ . Velocity is assumed to vary directly with  $z$ , the expected rate of change of outlays or nominal output. All of this is summarized by:

$$(1) \quad V \equiv \frac{1}{\phi_d},$$

$$(2) \quad \frac{dV}{dz} = -\frac{1}{\phi_d^2} \cdot \frac{d\phi_d}{dz} > 0$$

The expected rate of change of nominal output is calculated as a distributed lag on

of outlays which in turn are used to estimate the demand function for money, and hence velocity.

past actual rates,  $x$ , according to (Allais, equations (2.40), (2.28))

$$(3) \quad z(t) = \chi' \int_{-\infty}^t x(\tau) \exp \left[ - \int_{\tau}^t \chi(u) du \right] d\tau$$

The time variability of the distributed lag in (3) is a result of the fact that the weight attached to any observation  $t-\tau$  periods back is a function of the integral of  $\chi$  over the interval  $t-\tau$ . Since  $\chi$  varies with  $t$  this integral is not a function only of the length of the interval. This contrasts with the familiar time-invariant distributed lag used by Cagan (1956):<sup>3</sup>

$$(4) \quad z_1(t) = \chi' \int_{-\infty}^t x(\tau) \exp[-\chi'[t-\tau]] d\tau$$

It is possible to write the Allais distributed lag in terms of this more familiar Cagan one as

$$(5) \quad z(t) = \chi' \int_{-\infty}^t \chi(\tau) \exp[-\chi'[t-\tau]] \cdot \left\{ \exp \left[ \chi' \int_{\tau}^t -v(u) du \right] \right\} d\tau,$$

where

$$(5') \quad v = \frac{\chi - \chi'}{\chi'}$$

In other words, the Allais distributed lag can be thought of as one in which the Cagan

<sup>3</sup> It is possible, as a piece of simple mathematics, to choose a transformation of  $t$  that leads to expressing (3) in a time-invariant form like (4); in particular define  $\tau'$  by

$$\chi'[t-\tau'] = \int_{\tau'}^t \chi(u) du$$

Allais uses this result to construct a "shadow world" in which rates of transformation in terms of the transformed time scale are constant. (Borrowing one of the titles from the work of the novelist Ross MacDonald, we might call this world *The Far Side of the Dollar*.) Most commentators on Allais' work seem to be confused by the relationships between psychological and actual time (see Allais (1969b, 1970)). But whatever its intuitive appeal, the psychological time scale has one important application: it is used to specify a priori the values of some of the parameters of the model (see Allais 1966, pp. 1135-37).

distributed lag is modified, that modification being a function of  $v$ . The definition of  $\chi$  which Allais uses (1966, equation (2.25)) is

$$(6) \quad \frac{\chi}{\chi'} = \frac{\phi_0}{\phi_d}$$

where  $\phi_0$  is the value of  $\phi_d$  at  $z=0$ . Hence we have that

$$(7) \quad v = \frac{\chi - \chi'}{\chi'} = \frac{\phi_0 - \phi_d}{\phi_d} = \frac{V - V_0}{V_0}$$

The "perturbation element" represented by the second exponential in equation (5), therefore, is a function of what we might call *unit normal velocity*, and which we have denoted by  $v$ . If we expand  $v$  in a Taylor's series about  $z=0$ , we obtain

$$(8) \quad v = \frac{dV_0}{dz} \cdot \frac{1}{V_0} \cdot z + \frac{1}{2} \frac{d^2V_0}{dz^2} \cdot \frac{1}{V_0} \cdot z^2 + \dots = \epsilon + v(z), \quad z \rightarrow 0$$

where  $\epsilon$  is the elasticity of  $V$  with respect to  $z$ , evaluated at  $z=0$ . In the case in which  $z$  is the expected rate of change of prices, for example,  $\epsilon$  is the absolute value of the interest elasticity of the demand for money. Except in the extremes of hyperinflation the second and higher moments of  $v$  should be small enough that we can approximate (5) to a first-order term as

$$(9) \quad z(t) = \chi' \int_{-\infty}^t x(\tau) \exp[-\chi'[t-\tau]] \cdot \left\{ 1 - \chi' \int_{\tau}^t v(u) du \right\} d\tau$$

With the transformations  $u' = u - t$ , and  $s = t - \tau$ , this can be written as

$$(10) \quad z(t) = z_1(t) - \chi' \int_0^{\infty} x(t-s) \cdot \left\{ \int_{-s}^0 v(t+u') du' \right\} \exp[-\chi's] ds$$

Integrating by parts the inner integral of the second term  $N$  times, we obtain<sup>4</sup>

$$(11) \quad z(t) = z_1(t) - \chi' \int_0^\infty x(t-s) \cdot \left\{ \sum_{n=1}^N D^{n-1}v(t-s) \frac{s^n}{n!} + R_N \right\} \cdot \exp[-\chi's] ds$$

where

$$(11') \quad D^n = \frac{d^n(\cdot)}{dt^n}$$

To a first-order approximation, therefore, the difference between the Allais and Cagan estimates of  $z$  consists of weighted sums of products involving  $x$  and  $v$  and its derivatives. The shape of the distributed lag attached to each product is specified by the order of the derivative of  $v$ . For the  $j-1$  derivative of  $v$  the distributed lag function is given by

$$(12) \quad W_j(t) = \frac{t^j}{j!} \exp[-\chi't]$$

The sum of weights, the mean and the mode of the distribution are given respectively by:

$$(13) \quad S_j = \int_0^\infty W_j(t) dt = \frac{1}{\chi'^j},$$

$$(14) \quad \bar{t}_j = \frac{\int_0^\infty t W_j(t) dt}{S_j} = \frac{(j+1)}{\chi'},$$

$$(15) \quad \bar{t}_j^m = \frac{j}{\chi'}$$

<sup>4</sup> For the remainder,  $R_N$ , we have that

$$\begin{aligned} |R_N| &= \int_{-\infty}^0 |D^N v| \frac{u^N}{N!} du \\ &\leq \sup_u |D^N v| \int_{-\infty}^0 \frac{u^N}{N!} du \\ &\leq \sup_u |D^N v| \frac{s^{N+1}}{(N+1)!} \end{aligned}$$

and hence

$$\lim_{N \rightarrow \infty} |R_N| = 0 \quad \text{for all } s$$

The maximums of these distributed lags do not occur at  $t=0$  as is the case for the Cagan distributed lag. As can be seen from (15), they are placed further back in time the higher the order of the derivative of  $v$ . The same is true for the mean lags, given by (14); note, however, that for all  $j$  the *normalized* mean lag,  $\bar{t}_j - \bar{t}_j^m$ , equals  $1/\chi'$  which is the Cagan mean lag.

Allais (1966, equation 2.39) specifies the demand function for money as

$$(16) \quad \frac{\phi_d}{\phi_d} = [1 + b][1 + b \exp[\alpha Z]]^{-1}$$

where  $Z = z/\chi'$ . Writing this in terms of unit normal velocity,  $v$ , we have that

$$(17) \quad \begin{aligned} v &= \frac{V - V_0}{V} \\ &= \frac{\phi_0}{\phi_d} - 1 \\ v &= \frac{b}{1+b} + \frac{b}{1+b} [\exp[\alpha z/\chi']] \end{aligned}$$

To a first-order approximation, therefore, we have that

$$(18) \quad \begin{aligned} v &= \frac{2b}{1+b} + \frac{b}{1+b} \cdot \frac{\alpha}{\chi'} z \\ &= b_0 + b_1 z \end{aligned}$$

Substituting for  $z$  from (11) after dropping  $R_N$  we have for the estimated velocity function:

$$(19) \quad \begin{aligned} v(t) &= b_0 + b_1 z_1(t) - b_1 \chi' \int_0^\infty x(t-s) \\ &\quad \cdot \left\{ \sum_{n=1}^N D^{n-1}v(t-s) \frac{s^n}{n!} \exp[-\chi's] \right\} ds \end{aligned}$$

Hence the Allais method of estimating velocity comes down to estimating it as a distributed lag in itself and its derivatives. The weighting schemes of these distributed lags are concentrated further back in time the higher the order of the derivative of  $v$ . In terms of relating  $v$  to its own past behavior, this is not unreasonable. For example, since

a change in the second derivative takes longer to make itself felt on the level of  $v$  than does a change in the first derivative, we would want the distributed lag on the first to give more weight to more distant observations. Or to put the argument in terms of turning points: since the second derivative in  $v$  peaks before the first the weighting patterns should be ordered on the time domain to preserve this timing relationship, i.e., to be concentrated further back in time the higher the order of the derivative.

Of course  $x$  also appears in these distributed lags. But the variation across the terms in (19) is in the order of the derivative of  $v$ . In other words, the sum of terms representing the difference between the Allais and Cagan estimates is a series expansion in  $v$ , not  $x$ . And as we noted above, the shapes of the exponential weighting patterns are tailored to catch the timing relationships between  $v$  and its derivatives.

The  $x$  series can be thought of as part of the set of weighting coefficients in the distributed lag on velocity. It is their presence in (15) which makes the Allais distributed lag a variable-time one. Since the short-run behavior of velocity is pro-cyclical, the rates of change of nominal output and velocity are positively correlated. Hence the configuration of  $x$ -weights will tend to preserve the past variation in velocity. It may also add a delay to the impact of the past variation in the derivatives of velocity. For example, if velocity peaks before the rate of change of nominal output, the full impact of this turnaround will not be felt in the calculation of  $z$  until the weighting coefficients (the rates of change of nominal output) have also peaked.

Moreover the rates of change of nominal output are a relatively smooth series compared to the behavior of velocity. Again, therefore, the use of the  $x$ 's as weights should preserve the past configuration of velocity in the estimate of  $z$ , and hence in the current estimate of velocity.<sup>5</sup>

For all these reasons, the Allais procedure is approximately that of regressing velocity on a smoothed version of itself. This implies that the question of whether to use rates of change of outlays or of prices has nothing to do with which is the correct argument in the demand function for money. Rather the Allais procedure chooses only on the basis of which is a better smoothing function in estimating velocity from its past values.

#### REFERENCES

- M. Allais, "A Restatement of the Quantity Theory of Money," *Amer. Econ. Rev.*, Dec. 1966, 56, 1123-57.
- , (1969a) "Growth and Inflation," *J. Money, Credit, Banking*, Aug. 1969, 1, 355-426.
- , (1969b) "Growth and Inflation: A Reply to the Observations of the Discussants," *J. Money, Credit, Banking*, Aug. 1969, 1, 441-62.
- , "Allais' Restatement of the Quantity Theory: Reply," *Amer. Econ. Rev.*, June 1970, 60, 447-56.
- P. Cagan, "Allais' Monetary Theory: Interpretation and Comment," *J. Money, Credit, Banking*, Aug. 1969, 1, 425-32.
- , "The Monetary Dynamics of Hyperinflation," in M. Friedman, ed., *Studies in the Quantity Theory of Money*, Chicago 1956, 25-117.
- M. Darby, "Allais' Restatement of the Quantity Theory: Comment," *Amer. Econ. Rev.*, June 1970, 60, 444-46.
- H. G. Johnson, "Monetary Theory and Policy," *Amer. Econ. Rev.*, June 1952, 52, 344-45.
- A. Papoulis, *The Fourier Integral and Its Applications*, New York 1962.

<sup>5</sup> If the  $x$  series has sharply bounded derivatives, its frequency-domain representation given by its Fourier Transform will have its moments concentrated around zero. This means that the Fourier Transform of  $z$  given by (15) will be much like that for velocity. But since  $z$  is used to estimate velocity we should not be surprised that the Allais estimates are very good approximations (see Papoulis, pp. 32-34).

# The Phillips Curve and the Distribution of Unemployment

By A. G. HINES\*

The point of departure of this paper is the article by G. C. Archibald which appeared in a recent issue of this *Review*. That article contained some rather widespread views concerning the theory of wage inflation which can be legitimately questioned.

Archibald's main proposition may be summarized as follows: "Some of the explanatory variables appearing in fitted wage change equations . . . are *prima facie* inconsistent with the simple excess demand model" (p. 125).

According to this model,  $dW/W = f(X)$  where  $dW/W$  is the proportional rate of change of an index of money wage rates and  $X$  is the level of excess demand for labor expressed as a proportion of the labor force. Excess demand is linked via a transformation function to the level of unemployment. Hence,  $dW/W = g(U)$ . Archibald continues thus:

It follows from elementary price theory that if  $dW/W = f(X)$ , we shall have  $f'(0) = 0$ . Thus we should not find any variable such as prices or productivity (the rate of change of unemployment or union militancy) which affects either the demand curve or the supply curve, or both, apparently exercising an independent effect on  $dW/W$ :  $X$  should be the only argument of the function, the slope of which may, however, be affected by extraneous variables . . .

[p. 125, words in parentheses added]

Thus, according to Archibald, all variables other than the level of unemployment which enter wage adjustment equations are 'intruders.'

We shall argue that such variables are not intruders and that they can be shown on a priori grounds to have a place in the wage equation.

\* University of Durham. I am indebted to G. C. Archibald, J. D. Hey and N. Rau for comments on an earlier draft. Hey is joint author of Section III.B. of this paper.

## I.

Consider first the rate of change of prices ( $dP/P$ ). Contrary to Archibald's statement, it is not the case that elementary price theory postulates  $dW/W = f(X)$ ,  $0 = f(0)$ . As is well known, price theory postulates a relationship between the rate of change of *real* wage rates and the level of excess demand.

We have as equilibrium relationships

$$(1) \quad N^d - N^s = f(w | Y) - g(w | Z) = 0$$

where  $Y$  and  $Z$  are vectors of exogeneous variables the elements of which are parameters of shift of the demand and supply functions and  $w$  is the real wage rate ( $W/P$ ). The neoclassical dynamic adjustment hypothesis is<sup>1</sup>

$$(2) \quad \begin{aligned} dw/w &= \lambda[(N^d - N^s)/N^s] = \lambda(X) \\ &= \lambda h(w | Y, Z) \end{aligned}$$

Suppose following Phillips that there is a stable non-linear transformation between excess demand and the proportion of the labor force which is unemployed ( $U$ ), i.e.,

$$(3) \quad X = \theta(U)$$

$\theta$  being such that when  $X \rightarrow 0$ ,  $U \rightarrow a$ ,  $a > 0$  and when  $X \rightarrow \infty$ ,  $U \rightarrow 0$ . Then

$$(4) \quad dw/w = \theta(U)$$

Now,

$$(5) \quad dw/w = dW/W - dP/P$$

Therefore,

$$(6) \quad dW/W = \theta(U) + \beta dP/P, \quad \beta = 1$$

We therefore see that if it is anticipated, the

<sup>1</sup> The assumption which is customary in this connection, that  $dw/w$  is proportional to the level of excess demand, is strictly speaking ad hoc since price theory does not specify the precise form of the function. Some of the papers in Phelps et al. are devoted to an analysis of this problem.

rate of change of prices affects the rate of change of money wage rates quite independently of the level of excess demand as measured by the level of unemployment or any other such variable. Hence, if we adapt the argument of Milton Friedman, Edmund Phelps and others,  $dW/W$  depends on  $U$  alone only if the expected rate of price change is constant whatever is the actual rate of change of prices. In general, we would not expect this to be the case and hence  $dP/P$  is a legitimate variable in the wage equation. Moreover, if following Friedman we assume that the expected rate of inflation is proportional to the actual rate of inflation, the factor of proportionality being unity but that expectations take time to become fully adapted to the actual rate of inflation, the steady state solution to the Phillips curve yields a vertical line in the  $dW/W, U$  plane.<sup>2</sup> Now Archibald, apparently aware of the Friedman argument, merely notes that changing expectations shift the curve and still continues to regard  $dP/P$  as an intruder.

## II.

Consider next the variables which enter the vectors  $Y$  and  $Z$  such as the rate of change of productivity, and the rate of change of an index of union militancy, etc. Archibald's view that their presence in the wage equation is illegitimate is based on the same argument as is used to exclude  $dP/P$ , namely, that the effect of any such variables must show up in excess demand (as measured by the level of unemployment). Now the argument which is due to Richard Lipsey is undoubtedly correct. However, there is a question concerning the inference which is

<sup>2</sup> Suppose that the adaptive expectations hypothesis is appropriate to the formulation of expectations about price changes. Then, using  $p$  to denote  $dP/P$  and  $p^*$  to denote  $(dP/P)^*$ ,  $p_t^* = p_{t-1}^* + \psi(p_{t-1} - p_{t-1}^*) = \dots = \sum_{i=0}^{\infty} \psi(1-\psi)^i p_{t-1-i}$ . In the case of unit elasticity of expectations,  $\psi = 1$  and  $dW/W = f(X) + \beta p_{t-1}$ ,  $\beta = 1$  and  $p_t = p_{t-1}$  all  $t$ . The hypothesis  $\beta = 1$  is then directly testable. The case  $\psi = 0$  corresponds to the usual statement of the Phillips curve. In the more usual case,  $0 < \psi < 1$  and  $p^*$  is some weighted function of all past values of  $p_t$ . A  $p^*$  series must then be constructed in order to test the hypothesis  $\beta = 1$ . Solow has tested this hypothesis for U.K. and U.S. data for a wide range of values of  $\psi$  and found that invariably  $0 < \beta < 1$ .

to be drawn from it. For, suppose that  $U$  is a suitable proxy for the level of excess demand. Then the argument can be seen to imply that if in the wage adjustment equation we were to specify all the variables which are parameters of shift of the demand and the supply equations or are proxies for such parameters of shift, the level of unemployment or for that matter any other proxy for excess demand would become redundant. In other words, we can either have

$$(7) \quad dw/w = \lambda(X) = \theta(U)$$

or

$$(8) \quad dw/w = \lambda h(w | Y, Z)$$

We can also have

$$(9) \quad dw/w = f(U | Y, Z)$$

if we assume that given the inclusion of the elements of  $Y$  and  $Z$ ,  $U$  is a proxy for the levels of  $X$  which are associated with the adjustment of  $w$  to its equilibrium level as of given demand and supply curves. But, we cannot have

$$(10) \quad dw/w = \lambda h(w | Y, Z, U)$$

Which of the admissible alternative formulations we choose to have depends on the problem in hand. Thus we might assume a stable relationship between  $dW/W$  and  $U$  (given  $dP/P$ ) if we simply wish to predict  $dW/W$  conditional on  $U$ . However, we might wish to investigate the causes of inflation in order to be able to say something about the relative importance of the various *impulse* factors in the inflationary process over some given period. (The writer hazards the guess that it is this interest in the causes of inflation, as evidenced by the demand-pull cost-push debate, which has stimulated much of the work of the past fifteen or so years.) But, as Lipsey pointed out, knowledge of the shape and position of the adjustment function  $dW/W = f(U)$  does not enable us to distinguish between causes of disequilibrium. Our point is that we do wish to know something about the causes of inflation. And, if we attempt to obtain this information within the framework of neoclassical theory, we

require an explicit specification of the wage equation in terms of the variables which constitute the parameters of shift of the supply and the demand functions.

### III.

Finally, consider the rate of change of unemployment ( $dU/U$ ), the intruder which together with the trade union variable Archibald is most keen to repulse. Two matters require discussion here: the role of  $dU/U$  in the wage equation and the sign of the dispersion effect  $\partial(dW/W)/\partial\sigma_u^2$ .

A. It can be shown that  $dU/U$  may enter the wage equation with  $U$  as a joint proxy for excess demand.

Now

$$(11) \quad N^* = E + U$$

and

$$N^d = E + V$$

where  $E$  is the number of workers in employment,  $V$  is the number of vacancies, and  $U$  is the number of unemployed workers.<sup>3</sup>

$$(12) \quad \begin{aligned} X &= (N^d - N^*)/N^* = (V - U)/L, \\ L &= E + U \end{aligned}$$

If  $U/L$  is to be invariably a sufficient proxy for  $X$  to the exclusion of other variables, not only must  $X$  be constant when  $U$  is constant, and vary when  $U$  is varying, but it must not be the case that  $U$  systematically over- or understates the level of  $X$ . If we differentiate  $X$  with respect to time:

$$(13) \quad \begin{aligned} dX/dt &= (dV/dt - dU/dt)/L \\ &- (V - U)(dL/dt)/L^2 \end{aligned}$$

Hence, as of a given labor force,  $X$  can vary when  $U$  is constant if  $V$  varies independently of  $U$ . Hence, if  $U/L$  is to be a sufficient proxy for  $X$  when  $U/L$  and  $X$  are both varying,  $V/L$  must be constant. If  $dX/dt = 0$  and  $dL/dt = 0$ ,  $dV/dt - dU/dt = 0$ . Thus  $X$  can be

constant while  $U$  is varying if  $dV/dt = dU/dt$ . These possibilities are ruled out since over the cycle, we expect  $V$  to be positively related to  $X$ ,  $U$  to be negatively related to  $X$ , and  $V$  and  $U$  to be inversely related.

In a frictionless labor market, equilibrium would occur at zero vacancies and zero unemployment and  $X = V/L$  for  $X > 0$  and  $X = -U/L$  for  $X < 0$ . However, it is usually assumed that for various reasons—for example it takes time for workers to move between jobs, and a vacancy may exist in one place and the corresponding unemployed worker may be in another—zero excess demand is associated with positive unemployment and positive unfilled vacancies. This being the case  $U \rightarrow 0$  as  $X \rightarrow \infty$ . Similarly  $V \rightarrow 0$  as  $X \rightarrow -\infty$ . Since  $dV/dX > 0$  and  $dU/dX < 0$ , whenever  $dX/dt \neq 0$ ,  $dV/dt \neq dU/dt$ , the condition which is necessary if  $X$  is to vary whenever  $U$  varies. Specifically  $dV/dt > dU/dt$  when  $X$  is rising and  $dV/dt < dU/dt$  when  $X$  is falling so that  $|dX/dt| > |dU/dt|$  whenever  $X_t \neq X_{t-1}$ . Since  $X = (V - U)/L$ , this implies that  $U$  will understate  $X$  when  $X$  is rising ( $U$  is falling) and overstate  $X$  when  $X$  is falling ( $U$  is rising) for any given level of  $U$ . But this means that if employers bid for labor on the basis of the level of excess demand, they will bid more for labor when  $U$  is falling than when it is rising for any given level of  $U$ . To the extent that such bidding is successful, it reduces both  $U$  and  $V$ . However, given that the relationship between  $V$  and  $U$  is non-linear it reduces  $U$  by less than it reduces  $V$ . Hence,  $dW/W$  will be greater when  $U$  is falling than when it is rising for any given level of  $U$ . Moreover, it can be shown (see Hines, 1971) that the misstatement of  $X$  by  $U$  varies in a manner which is determined by the non-linearity in the relationship between  $V$  and  $X$ ,  $U$  and  $X$ , as well as by the pattern of cyclical variations in  $X$ . Thus if vacancy statistics are not available so that  $X$  is not directly measurable, in nonstationary situations,  $dU/U$  is a valid proxy with  $U/L$  for the level of excess demand. The same conclusion which is established by Phelps et al. in a different manner is noted by Archibald. However, he does not draw the obvious con-

<sup>3</sup> Except in the remainder of this paragraph where it refers to the number of unemployed workers,  $U$  represents the proportion of the labor force which is unemployed ( $U/L$ ).

clusion for his proposition that  $dU/U$  is an intruder. For it is true that if we are simply interested in estimating the parameters of the adjustment function,  $dW/W = f_0(X)$  given  $dP/P$ , all other variables are legitimately excluded. But it does not follow that if we now choose  $U$  as a proxy for  $X$  that  $dW/W = f_1(U)$  to the exclusion of  $dU/U$ .

B. Archibald shows that a sufficient condition for the dispersion effect to be positive in the case of two sectors is that the slope of the Phillips curve in sector 1 should be greater than the slope of the Phillips curve in sector 2, given  $U_1 > U > U_2$ . This may be directly established as follows.

$$(14) \quad U = \alpha_1 U_1 + \alpha_2 U_2$$

where  $\alpha_1$  is the fraction of the labor force in sector 1 and  $\alpha_2 = 1 - \alpha_1$ .

$$(15) \quad dW/W = \alpha_1 f_1(U_1) + \alpha_2 f_2(U_2)$$

where  $f_1$  and  $f_2$  are the Phillips curves in sectors 1 and 2. From (14),  $dU = \alpha_1 dU_1 + \alpha_2 dU_2$ , and so for constant  $U$  we require  $dU = 0$ , i.e.,  $dU_1 = (-\alpha_2/\alpha_1)dU_2$ . From (15),

$$d(dW/W) = \alpha_1 f'_1(U_1)dU_1 + \alpha_2 f'_2(U_2)dU_2$$

Hence,

$$(16) \quad d(dW/W) = -\alpha_2 [f'_1(U_1) - f'_2(U_2)]dU_2$$

for constant  $U$ .

Now,  $\sigma_u^2 = \alpha_1(U_1 - U)^2 + \alpha_2(U_2 - U)^2 = \alpha_2/\alpha_1(U_2 - U)^2$ , using (14) and the condition  $\alpha_1 + \alpha_2 = 1$ ; and so,  $d\sigma_u^2/dU_2 = 2\alpha_2/\alpha_1(U_2 - U) = 2\alpha_2(U_2 - U_1)$  for constant  $U$ . Hence

$$(17) \quad \frac{d(dW/W)}{d\sigma_u^2} = \frac{d(dW/W)}{dU_2} \frac{dU_2}{d\sigma_u^2} = \frac{f'_1(U_1) - f'_2(U_2)}{2(U_1 - U_2)}$$

And, in order that  $d(dW/W)/d\sigma_u^2 > 0$  for constant  $U$  and  $U_1 > U > U_2$ , we require

$$(18) \quad f'_1(U_1) > f'_2(U_2)$$

which is basically the result which Archibald obtained.

However, there is no a priori reason why since the nineteenth century the sector (industry as well as region) whose Phillips curve has the larger slope at all points should invariably have the greater unemployment. Moreover, the empirical evidence does not show this to be the case. For completeness, we should therefore consider the possibilities  $U_1 = U = U_2$  and  $U_1 < U < U_2$ .

Now starting from  $U_1 = U = U_2$ , suppose that for constant  $U$ , the variance is increased by increasing  $U_1$  and decreasing  $U_2$ . Then in order that  $d(dW/W)/d\sigma_u^2 > 0$ , we require

$$(19) \quad f'_1(U) \geq f'_2(U)$$

Conversely, if the variance is increased by increasing  $U_2$  and decreasing  $U_1$  for constant  $U$ , in order that the dispersion effect be positive, we require, by symmetry, that

$$(20) \quad f'_1(U) \leq f'_2(U)$$

But (19) and (20) taken together, imply that in order that  $d(dW/W)/d\sigma_u^2 > 0$ .

$$(21) \quad f'_1(U) = f'_2(U)$$

for any  $U$ , i.e., the Phillips curve in sectors 1 and 2 must be *parallel*. If the dispersion effect is to be positive while (21) also holds the Phillips curve in each sector must be *convex* to the origin.

We note further that  $d(dW/W)/d\sigma_u^2$  is not in general constant since it depends on the chosen values of  $U_1$  and  $U_2$ . It will only be constant if  $f'_1(U_1) - f'_2(U_2) = k(U_1 - U_2)$ , i.e., if

$$(22) \quad \begin{aligned} f_1(U_1) &= a_1 + b_1 U_1 + \frac{k}{2} U_1^2 \\ f_2(U_2) &= a_2 + b_2 U_2 + \frac{k}{2} U_2^2, \end{aligned}$$

for some  $b$ , and  $k > 0$

and from equation (15),

$$(23) \quad dW/W = A + BU + C\sigma_u^2 + DU^2$$

where  $A = (\alpha_1 a_1 + \alpha_2 a_2)$ ,  $B = b_1 = b_2$ ,  $C = D = k/2$ . For, recalling equation (21) which ensures that the ranking of sectors by the distribution of unemployment can vary, we see that for  $d(dW/W)/d\sigma_u^2 = \text{constant}$ ,

$$\frac{f'_1(U_1) - f'_2(U_2)}{U_1 - U_2} = \frac{f'(U_1) - f'(U_2)}{U_1 - U_2} = f''(\xi)$$

for some  $\xi$  in  $(U_1, U_2)$  where  $f' \equiv f'_1 \equiv f'_2$ , i.e.,  $f'_1(\xi) = f'_2(\xi) = \text{constant}$  for all  $\xi$ .

In other words, as in equation (23), the parallel convex Phillips curves must have a *constant second derivative*. But equation (22) bears no resemblance to the functional forms which have been used to describe the Phillips curve. Hence contrary to Archibald's procedure in his empirical work,  $\sigma_u^2$  cannot be included linearly in an equation of the form  $dW/W = f(U, \sigma_u^2)$  since strictly speaking  $\partial f / \partial \sigma_u^2$  is not a constant and may even vary more than  $\sigma_u^2$  itself.

#### IV.

Turning to the empirical evidence, Archibald reports that for postwar U.K. data not only is  $dU/U$  insignificant in an equation which excludes  $\sigma_u^2$  but also that  $\sigma_u^2$  is significant (whereas  $dU/U$  is not) in the absence as well as in the presence of  $dU/U$  in the wage equation.<sup>4</sup> It might also be instructive to look at the results which have been obtained from equations which have been fitted directly to disaggregated data. Keith

<sup>4</sup> Using the  $t$ -statistic as a criterion of goodness of fit, Archibald's results are firmer for the equations in which  $\sigma_u^2$  is measured across industries. Nevertheless  $\sigma_u^2$  makes a significant contribution to the proportion of explained variation in  $dW/W$  in both models. A. Thirlwall whose work covers a similar period finds that  $\sigma_u^2$  contributes significantly in the industrial model but not in the regional model. In commenting on his findings Archibald writes as follows:

On this evidence, we might conclude that there is a Phillips curve, and that a reduction in the regional dispersion of unemployment in the U.K. would move the fiscal policy frontier in a favourable direction. [p. 129]

In contrast, Thirlwall concludes thus:

We are forced to the conclusion again, therefore that the dispersion of regional rates of unemployment does not appear to have exerted a significant independent influence on wage rate inflation in the postwar period that is independent of variations in the aggregate state of demand and the dispersion of industry rates of unemployment. This is an important conclusion in view of statements often made (but unsupported by positive evidence) to the effect that if regional disequilibrium in the economy could be diminished the pace of wage inflation nationally could be curbed. [p. 76]

Cowling and David Metcalf fit an equation of the type  $dW/W = f(U, dU/U)$ , in which the dependent variable is the rate of change of an index of earnings to pooled regional U.K. data over the period 1960-65. They find that the coefficient of  $dU/U$  is significant but that of  $U$  is not significantly different from zero by conventional standards. Metcalf fits similar equations separately to postwar data for ten U.K. regions. He reports significant coefficients on  $U$  and  $dU/U$ . However, the coefficients of  $U$  do *not* have the expected sign. In my 1969 article, I reported that when a similar model was fitted to U.K. data for twelve industries over the period 1948-62,  $U$  had the expected sign in all but one case and was significant in three cases whereas  $dU/U$  was insignificant and had the expected sign in one case only. The evidence from regional and industry data are not entirely consistent, particularly concerning the significance of  $dU/U$ . Nevertheless both sets of data do raise doubts about the existence of adjustment functions of the Phillips type at these levels of disaggregation. Moreover, contrary to Lipsey's aggregation hypothesis on which Archibald's arguments are erected,  $\sigma_u^2$  and  $dU/U$  are not correlated. Further, the data do not show that the slopes of the sector Phillips curves are identical, a condition which must be met if the sign of the dispersion effect is to be unambiguously positive. (In some cases, sector Phillips curves have been found to be upward sloping.) Hence, on the available evidence it does not appear as if  $dU/U$  is a proxy for  $\sigma_u^2$  in the wage equation, and since we have shown that the sign of the dispersion effect is not in general unambiguous, we must then ask if there is an alternative rationalisation of the observed correlation between  $\sigma_u^2$  and  $dW/W$ . In my 1971 unpublished mimeograph I attempt an answer in terms of the transfer mechanism hypothesis, according to which groups in the bargaining process in different sectors are interrelated in such a way that wage changes in lead sectors, whatever their cause, are transmitted to following sectors.

#### V.

I have examined and rejected the proposi-

tion that on a priori grounds, all variables other than the level of unemployment are intruders in any wage adjustment equation which contains the level of unemployment as an explanatory variable. The expected rate of change of prices (for which the actual rate of change is a proxy) finds a place because the neoclassical theory which is usually taken to be the basis of such equations specifies relationships in real wages. In nonstationary situations the rate of change of unemployment is validly a joint proxy with the level of unemployment for the level of excess demand in the absence of vacancy statistics. Moreover, we can either operate with a wage adjustment function expressed in terms of the excess demand for labor, choosing suitable proxies for excess demand, or operate in terms of explanatory variables such as the rate of change of productivity or the rate of change of unionization which are parameters of shift of the demand and supply equations. It is only in the latter formulation that given this framework, we can hope to learn something about the impulse factors in the inflationary process.<sup>5</sup>

I have examined the conditions under which a positive relationship between the rate of change of money wage rates and the sectoral distribution of unemployment can be deduced from the Lipsey aggregation hypothesis as developed by Archibald. In the case of two sectors, the condition which is pretty stringent is that the slopes of the sector Phillips curves be identical, each curve being convex with a constant second derivative. The results which we have obtained for the two sector case, can be generalized to the case of  $n$  sectors,  $n > 2$ .

One conclusion of this paper is that neither the theory of the Phillips curve nor the available empirical evidence unambiguously indicate that a policy of reducing the sectoral inequality in the distribution of unemploy-

ment—a policy which is to be encouraged on many grounds and in its own right—will reduce the aggregate rate of change of money wage rates.

## REFERENCES

- G. C. Archibald, "The Phillips Curve and the Distribution of Unemployment," *Amer. Econ. Rev. Proc.*, May 1969, 59, 124-34.
- K. Cowling and D. Metcalf, "Wage Unemployment Relations: A Regional Analysis for the U.K., 1960-1965," *Oxford Inst. Econ. Statist. Bull.*, Feb. 1967, 29, 31-39.
- M. Friedman, "The Role of Monetary Policy," *Amer. Econ. Rev. Proc.*, May 1968, 28, 1-17.
- A. G. Hines, "Trade Unions and Wage Inflation in the United Kingdom 1893-1961," *Rev. Econ. Stud.*, Oct. 1964, 31, 221-52.
- , "Wage Inflation in the United Kingdom, 1948-1962: A Disaggregated Study," *Econ. J.*, Mar. 1969, 79, 66-89.
- , "What Does the Phillips Curve Show?" mimeo. 1971.
- R. G. Lipsey, "The Relation Between Unemployment and the Rate of Change of Money Wage Rates in the United Kingdom, 1862-1957: A Further Analysis," *Economica*, Feb. 1960, 27, 1-31.
- D. Metcalf, "The Determinants of Earnings Changes: A Regional Analysis for the U.K. 1960-1968," *Int. Econ. Rev.*, June 1971, 12, 273-82.
- E. Phelps et al., *Microeconomic Foundations of Employment and Inflation Theory*, New York 1970.
- A. W. Phillips, "The Relation Between Unemployment and the Rate of Change of Money Wage Rates in the United Kingdom, 1861-1957," *Economica*, 1958, 25, 283-99.
- R. M. Solow, *Price Expectations and the Behaviour of the Price Level*, Manchester 1969.
- A. Thirlwall, "Demand Disequilibrium in the Labour Market and Wage Inflation in the United Kingdom," *Yorkshire Bull. Econ. Soc. Res.*, May 1969, 21, 66-76.

<sup>5</sup> However, in our view the attempt to rationalize the observed relationship between  $dW/W$  and  $U$  within this framework encounters some formidable if not insuperable objections which do not apply to at least one alternative framework. See Hines (1971).

# Uncertainty and the Evaluation of Public Investment Decisions: Comment

By E. J. MISHAN\*

Kenneth Arrow and Robert Lind propose to modify a familiar argument, which they associate with Jack Hirshleifer, to the effect that the use in public investment of a discount rate lower than the expected rate of return in the private sector can result in a displacement of private investment projects by public ones yielding lower returns.

Ignoring, for the moment, the crucial risk factor, this argument is valid in so far as public investment entails an equal reduction in current private investment. If, on the other hand, the public investment is to be financed wholly by a reduction in current consumption, there will be a gain to society even though the return expected on the public investment is lower than that on private investment *provided* that the former exceeds society's rate of time preference. Thus, a political constraint requiring that funds, if raised wholly by reducing current consumption, be confined to investment in public projects would warrant the use of society's rate of time preference as discount rate in evaluating their economic justifiability. Where funds so raised are not, however, constrained to public investments—so that they could as well be invested in the private sector, if it were deemed socially preferable to do so—the use of the return on private investment as discount rate is warranted.<sup>1</sup> For it ensures that no authorized funds—whether raised through reducing resources available to private consumption or private investment—are invested in the public sector when they can more profitably be invested in the private sector.

\* London School of Economics and American University.

<sup>1</sup> No such political constraint enters into the framework of assumptions within which Stephen Marglin and others analyze the opportunity cost of public investment; hence my proposal (1967) to employ, within that framework, the certain yield in the private sector as the appropriate rate of discount.

Under conditions of certainty, at least, we may conclude that the employment of the return on private investment as the discount rate is valid 1) if investible public funds are raised wholly by displacing private investment, or 2) if, however the funds are raised (whatever the combination, that is, of taxation and government borrowing), there is no restriction to their use in the private sector.<sup>2</sup>

## I

The introduction of risk-bearing into the analysis enables Arrow and Lind to require that another condition be met if the yield on private investment is to be used as the appropriate discount rate for public investment; namely, that the (subjective) cost of risk-bearing be the same for taxpayers as for private individual investors.

They argue that if benefits and costs are to be measured in terms of willingness to pay,<sup>3</sup> the costs of risk-bearing must be subtracted from the net benefits of investments in order to obtain a correct measure of their value to the recipients. According to their analysis, however, where the numbers of taxpayers are large, the risk borne by each in respect of any particular investment project

<sup>2</sup> In fact, so long as society's rate of time preference is below the certain yield in the private sector, *any* reduction of current consumption in favor of private investment is warranted. The only relevant question with respect to funds so raised is, then, whether they should be invested in the public sector at a rate of return no lower than that which prevails in the private sector or whether, instead, they should be invested in the private sector.

<sup>3</sup> More accurately, benefits and costs are to be measured by their compensating variations: individuals are willing to pay a maximum sum for benefits but require a minimum compensation for losses received. If the payments take on a positive sign and the losses a negative one then, granted relative prices are unchanged, a Kaldor-Hicks improvement is registered if the algebraic sum of compensating variations of all persons affected is positive.

is negligible. In contrast, the risk-bearing costs of a similar project to a limited number of private investors can be substantial. They conclude that it is not the government's pooling of investment projects so much as its spreading of the risk over a large number of taxpayers that would rationalize the ignoring of uncertainty in the evaluation of public investments. As a corollary, it follows that a public investment having an expected rate of return below that of private investments may yet be economically superior. For what is relevant in the comparison is not the expected rates of return *per se*, but the expected rates of return *net* of the costs of risk-bearing.

In drawing their conclusions, the authors invoke the Kaldor-Hicks criterion, though under the safeguarding assumption that relative prices remain unchanged. This is, of course, perfectly acceptable in this connection, since principles of resource allocation and benefit-cost analyses are ultimately founded upon such a criterion. Recall, however, that the intent of the Kaldor-Hicks criterion was to circumvent the distributional changes accompanying any economic reorganization. It posits costless transfers of goods, or income, and is met in moving from an economic arrangement I to II, if in the II position there is a distribution of the goods which could make none worse and some better off than they were in the I position.

On the assumption that, under conditions of certainty, society's rate of time preference is, say 5 percent, the undertaking of *either* the public investment or the private investment, in their example, is able to meet the Kaldor-Hicks criterion. The question, then, is to determine which of these two options is the better. The authors use an example to show that a decision to adopt the public investment option rather than the private investment one can make everyone yet better off.

To illustrate their argument, and my counter-argument which follows, I shall concoct a yet simpler example than theirs, one in which the social rate of time preference is zero and, for lack of a decent market for insurance against investment risks, private

investors subtract \$0.3 million from a yield of \$1.5 million that is expected in the following year, being the expected outcome of a current investment of exactly \$1.0 million. If, however, the government invests this same sum it could expect a return of \$1.25 in the following year.<sup>4</sup> Should the public rather than the private investment be undertaken in these circumstances? Yes, the authors would answer. For the private investors will be indifferent as between an expected return next year of \$1.5 million and certain receipt of \$1.2 million. The true opportunity cost of investing \$1.0 million in the public sector is, therefore, \$1.2 million (*not* \$1.5 million), and a discount rate of 20 percent (*not* 50 percent) should be employed there. Since the public investment effectively offers a certain \$1.25 million to taxpayers, by undertaking it the government can more than pay the opportunity loss of \$0.2 million suffered by private investors in foregoing the private investment, and yet have something left over for the taxpayers. Thus everyone can be made better off by adopting the public investment than he would have been if instead the private investment had been undertaken.<sup>5</sup> There appears, in fact, to be a potential net gain to society of \$0.05 millions from adopting the public rather than the private investment.

## II

The Arrow-Lind arrangement whereby private investors are compensated to forego their private investment for a present sum of \$1.2 million, on the view that it is in fact really only worth \$1.2 million to them as compared with the worth of the public investment of \$1.25 million to the taxpayer is not, however, the only line of argument possible. An alternative involves an arrangement in which the private investment option is adopted, and private investors are then compensated by \$1.2 million on trans-

<sup>4</sup> We can assume, further, that for both the private and the public project the variance is the same.

<sup>5</sup> Hirshliefer's recommendation that a direct subsidy to private investment be used in such cases is dismissed as involving simply a transfer payment without effectively reducing the private costs of risk-bearing.

ferring their claim to the taxpayers who value it at \$1.5 million. This is superior to the Arrow-Lind arrangement to the extent of \$0.25 million.

In more detail, the government, having originally decided to invest \$1 million in the public investment in expectation of paying in return \$1.25 million to the taxpayers, now changes its mind. It allows private investors to go ahead, and, after their doing so, transfers to them \$1.2 million in exchange for the expected return which is worth \$1.5 million to the taxpayers. Taxpayers alone, therefore, net \$0.3 million (instead of the \$0.25 million they were to have gained from the public investment) and are made, incidentally, better off by \$0.05 million as compared with the public investment.<sup>6</sup> But, in addition, private investors are made \$0.2 million better off than they themselves would have been had the choice been in favor of the public investment. The excess gain of private over the public investment option is, under this arrangement, not surprisingly the difference between \$1.5 million and \$1.25 million, or \$0.25 million. Finally, it should be noticed that the arrangement proposed above does not require to be implemented: actual transfers do not have to take place. For the Kaldor-Hicks criterion addresses itself strictly to *potential* improvements: hypothetical distributions only are envisaged.

There is, however, a feature that is peculiar to the preceding hypothetical Pareto improvement. It transpires that it is the very act of the transfer itself that enlarges the worth of the receipt for society from \$1.2 million to \$1.5 million. So that while it is literally true (a) that a hypothetical distribution of the output in the private investment situation could indeed make every one better off than he would have been in the public investment situation and, therefore, in moving from the public to the private investment option the Kaldor-Hicks criterion does appear to have been met, it is

also true (b) that, unless the transfer actually does take place, gains in the private investment situation do not exceed those in the public investment situation. As is so often the case in such matters then, the correct criterion would seem to depend upon the political and/or the administrative constraints that are assumed.

Allowing that in all cases at issue investment in the public sector entails foregoing the opportunity of spending in the private sector, the Arrow-Lind amendment appears to be valid if the public agency is permitted to use the funds allocated to it only to undertake specific investments. If, however, the public agency is permitted to use funds appropriated for investment purposes either to invest directly in the private sector, so availing itself fully of any actuarial rate of return prevailing in the private sector, or else to buy out private investment projects already undertaken—buying them out at (not less than) the capitalized value of the certainty-equivalent of their expected stream of benefits—the Arrow-Lind amendment does not apply. For under either of the latter options the opportunity rate of return open to the public funds will again be the full actuarial rate of return on private investment, which private rate of return can then be adopted as the appropriate rate of discount for public investment projects. Moreover, inasmuch as permitting public agencies these additional options of investing in, or buying from, the private sector promotes optimal use of investible funds, the economist should recommend them.

### III

There is finally—even under the political constraints favorable to the Arrow-Lind amendment—another argument in favor of regarding the unmodified expected (or actuarial) return on private investment as the proper opportunity yield of public investments. However equitable are the rules for distributing the national product, any society concerned with the growth of its material output would want to maximize the resulting returns on its investments over the future. Irrespective then of private inves-

<sup>6</sup> Assuming, as we do here, that under the public investment option, taxpayers were to receive the whole of the net gain of \$0.25 million. This assumption can, of course, be modified without changing the conclusion.

tors' subjective costs of risk-bearing, society might vote unanimously never to displace a private investment by a public one having lower expected yields. Society might vote this way in the knowledge that if instead it adopted the Arrow-Lind amendment it would, after the elapse of a number of years, be producing a national output smaller than it would have produced if it had continued to follow the alternative rule.

Thus even if *ex ante* real welfare is lower (inasmuch as the diswelfare suffered by risk-averse private investors has to be subtracted from the expected returns to private investment) the adoption of the alternative rule enables the community to attain a larger *ex post* output. And society might regard a larger *ex post* national output as preferable not only in its own right but also in its offering a larger output base from which a larger volume of investment will generate

further increases in national output—for future investment decisions depend *inter alia* not upon *ex ante* real income but upon *ex post* or realized income.

#### REFERENCES

- K. J. Arrow and R. C. Lind, "Uncertainty and the Evaluation of Public Investment Decisions," *Amer. Econ. Rev.*, June 1970, 60, 364-78.
- J. Hirshleifer, "Investment Decisions Under Uncertainty: Application of the State-Preference Approach," *Quart. J. Econ.*, May 1966, 80, 252-77.
- S. Marglin, "The Social Rate of Discount and the Optimal Rate of Investment," *Quart. J. Econ.*, Feb. 1963, 77, 95-111.
- E. J. Mishan, "Public Investment Criteria: Some Simplifying Suggestions," *J. Polit. Econ.*, Apr. 1967, 75, 139-46.

# Uncertainty and the Evaluation of Public Investment Decisions: Comment

By ROLAND N. MCKEAN AND JOHN H. MOORE\*

Using Pareto optimality (in the Hicks-Kaldor sense) as their criterion throughout, Kenneth Arrow and Robert Lind argue in the June 1970, issue of this *Review* that 1) for public investments the cost of risk-bearing should be regarded as zero because this cost is spread over a large number of persons; 2) consequently, public investment should displace private investment if the expected rate of return exceeds the expected return to private investment minus an adjustment for the cost of risk-bearing; 3) furthermore, project costs borne publicly or benefits accruing to government should be discounted at relatively low rates (because the cost of risk-bearing is low if spread among large numbers of persons), but project costs borne privately or benefits accruing to private individuals should be discounted at relatively high rates. Arrow and Lind are abstracting from other factors, e.g., externalities, public good characteristics, or ideological preferences for either state or private activity, that might also affect the choice between public and private investments.

We wish to emphasize anew the fundamental defect in any proof that a policy yields a Hicks-Kaldor improvement. We are not referring to the objection to the Hicks-Kaldor *criterion*—the fact that without actual compensation there will be a redistribution to which one may attach negative value. (Arrow and others have stressed that for this reason one cannot say that a Hicks-Kaldor change is a gain in welfare.)<sup>1</sup> We are referring rather to the fact that, without actual purchase of everyone's consent, one lacks *information* about whether

the gains exceed the cost, i.e., about whether it would in fact be possible to make some better off without making anyone worse off.<sup>2</sup> One may *judge* that a policy, such as public investment, would be a Hicks-Kaldor improvement because he *believes* the relevant tradeoffs in individuals' preference surfaces are such as to make the gains exceed the costs (as seen by each individual for himself). But he cannot show others that this is so: a Hicks-Kaldor improvement is by definition a change such that one can never demonstrate that it is a Hicks-Kaldor improvement!

The Arrow-Lind argument is in difficulty on this score because of the alternatives it considers, a public investment financed by taxes versus a private investment. In the latter case, individuals invest voluntarily, taking into account their marginal time preferences as well as their risk preferences. In a public investment financed by taxes, people are *forced* to invest. One has no observable data on whether all of these individuals would be willing to invest rather than consume, or data on how much they would have to be paid to invest voluntarily.<sup>3</sup> Some of them might much prefer to consume, given the circumstances assumed by Arrow and Lind: "At the margin, different

<sup>2</sup> See Harold Demsetz, pp. 67–68.

<sup>3</sup> Arrow and Lind acknowledge that spreading the risk over the taxpayers would not put investment in the hands of those persons having the least risk aversion. Our point is that it fails also to allow the investment to be made by those having the least unwillingness to forego consumption and to ascertain the cost of forcing taxpayers to exercise a particular marginal *time* preference.

Needless to say, such lack of information exists for any program financed by compulsory taxes. Jack Hirshleifer's proposal to subsidize private investment would involve the same difficulty except that he suggests having government *borrow* the funds, p. 270. Arrow and Lind, however, explicitly deal with spreading the risk over all *taxpayers*.

\* Professor and associate professor of economics at the University of Virginia, respectively. We wish to thank Roger P. Sherman for criticisms and suggestions.

<sup>1</sup> See Arrow, pp. 38–45. For other references, see Jerome Rothenberg, pp. 65–66.

individuals will have different rates of time and risk preference . . . " (p. 366),<sup>4</sup> because of imperfect insurance and capital markets.

While this lack of data exists whenever taxation is part of an alternative being evaluated, it does not make all cost-benefit analysis irrelevant. One might compare the Pareto-relevant effects of alternative government investments, taking the taxes as *given* and not counting them as part of the decisions being evaluated. If the cost-benefit comparisons were accurate, one could then identify investments that were Pareto optimal changes *given these constraints* (i.e., a suboptimization). In any event, all that we would say for cost-benefit analysis is that it may produce information that is pertinent to particular choices. But Arrow and Lind are not discussing the type of comparison noted above or saying merely that the information produced is relevant. They are discussing the comparison of a private investment with a public-investment-cum-taxes and identifying the public investment (in certain circumstances) as the preferred one. In this comparison the lack of information inherent in Hicks-Kaldor "improvements" does cause trouble.

But let us bypass this problem and also accept, for the moment, the Arrow-Lind argument that the total cost of risk-bearing is zero for government investments. While Arrow and Lind point out some implications, there are additional implications that are somewhat startling. If their proposition is correct, then *all* investments should be undertaken by government or *all* carried out privately, depending upon whether people are, on average, risk-aversers or risk-lovers.

Consider any investment and assume that people are risk-averse. Suppose the expected return was 5 percent or 10 percent or 15 percent. In every instance there would be some cost of risk-bearing if the investment were made privately but, according to Arrow and Lind, zero cost of risk-bearing if government

made the investment. It would *always* be more efficient, therefore, for government to build the shoe factory or the steel plant.<sup>5</sup>

On the other hand, suppose that people are risk-lovers. As Arrow and Lind point out, there would then be a negative cost, i.e., a positive gain, attached to risk-bearing, p. 373. In those circumstances, *all* investments should be private. The market discount rate would be below the riskless rate, and the government, since the negative costs of risk-bearing would be zero for public investment, should always discount at a *higher* rate than would be appropriate for private investors.

The final paragraphs of Arrow and Lind have a still more surprising implication. To see this, assume risk aversion again. As noted above, they state that project costs borne and benefits accrued privately should be discounted at a relatively high rate to reflect risk, while costs borne and benefits accrued publicly should be discounted at a lower rate. As a consequence, they say, "It is somewhat ironic that the practical implication of this analysis is that for the typical case where costs are borne publicly and benefits accrue privately, this procedure will qualify fewer projects than the procedure of using a higher rate to discount both benefits and costs" (p. 378). We think the implication is even more ironic if one considers the untypical but easily arranged case where the costs are borne privately and the benefits accrue publicly. (This could be arranged by taxing a few selected individuals to finance each project instead of spreading the costs over a large number of persons, and by producing benefits that go to large numbers of persons or having the beneficiaries pay the government for their benefits.) In these circumstances, the costs should be dis-

<sup>4</sup> They may have neglected the compulsory taxes because they compressed everything into one time period, in which case an investment means merely some net returns. But in real life the taxes come first and the returns later.

<sup>5</sup> One could bring in the possibility that expected returns might be lower with public ownership, but this did not figure in the original article, so we too put this possibility aside for purposes of this note. Also, as noted at the outset, we (like the authors) are abstracting from externalities, public good characteristics, ideological preferences, and other such additional variables that might affect the choice between public and private investment.

counted at a high rate reflecting risk, the benefits at a lower rate. Many additional investments would now appear to be worthwhile, and the riskier the project, the better it would appear to be! Government would choose the riskiest projects among those yielding the same nominal returns and the riskiest design of any one project. With the assumed risk aversion, does this really sound like Pareto optimality?

We do not believe, however, that any one should take either this Arrow-Lind case or ours seriously. After all in neither case are investors taking any risk at all. They are simply being forced to part with some money in exchange for an assured zero return *to them*, since in both cases the returns would go to a different set of persons from the investors.<sup>6</sup>

Where does all this leave the Arrow-Lind argument? As noted before, there is a fundamental criterion difficulty when a compulsory tax is a component of any alternative being evaluated. The reasonable way to view the argument, however, is presumably not to ask if their criterion is definitive, abstracting from all other considerations that might bear on decisions about public investment, but rather to ask if their risk-spreading argument is relevant, as one of several considerations, in making those investment choices. Even in this context, however, we are not convinced that much weight should be attached to the risk-spreading point. Some of the implications of the Arrow-

<sup>6</sup> It is generally hazardous to discount benefit and cost streams differently if the discounting is to reflect risk. When alternative ways of producing a specified physical capability are compared, researchers often discount the cost streams only. If the rate includes a cost of risk-bearing, the riskier design will have the advantage. The fallacy here is that the alternative designs do not in fact produce identical capabilities if there are differential degrees of risk.

Lind article—the super-sensitivity of its policy implications to the structure of risk preferences and the difficulty of finding out about that structure—raise doubts about the applicability of their argument. The fact that everything hinges on the number of taxpayers intensifies these doubts. After all, if the U.S. Defense Budget was spread over an infinite number of persons, the cost to individuals would approach zero.<sup>7</sup> With a finite number of taxpayers, however, the cost to individuals depends upon the absolute size of the total cost to be spread and upon the actual number of taxpayers. And in this finite world does the total cost sum back up to zero? As Arrow and Lind say: “The question necessarily arises as to how large  $n$  must be to justify proceeding as if the cost of publicly-borne risk is negligible. This question can be given no precise answer . . . ” (p. 373). This is correct. But eighty million [taxpayers] is a long way from infinity, and we are doubtful that the number is large enough to enable the United States to get very much for nothing.

#### REFERENCES

- K. J. Arrow, *Social Choice and Individual Values*, 2d ed., New York 1963.
- and R. C. Lind, “Uncertainty and the Evaluation of Public Investment Decisions,” *Amer. Econ. Rev.*, June 1970, 40, 364–78.
- H. Demsetz, “Some Aspects of Property Rights,” *J. Law Econ.*, Oct. 1966, 9, 61–70.
- J. Hirshleifer, “Investment Decision Under Uncertainty: Choice-Theoretic Approaches,” *Quart. J. Econ.*, Nov. 1965, 79, 509–36.
- J. Rothenberg, *The Measurement of Social Welfare*, Englewood Cliffs 1961.

<sup>7</sup> This statement is correct, but it might be noted that it is not an example of the proposition proved by Arrow and Lind.

# Uncertainty and the Evaluation of Public Investment Decisions: Comment

By ALAN NICHOLS\*

In the June 1970 issue of this *Review*, Kenneth Arrow and Robert Lind prove that "when the risks associated with a public investment are publicly borne, the total cost of risk-bearing is insignificant . . ." This result follows from the fact that the risk apportioned to any individual taxpayer from the adoption of a project is negligible, and, as they show, so is the sum of risks taken over all taxpayers. From this correct result they draw the unwarranted conclusions that "... therefore, the government should ignore uncertainty in evaluating public investments . . ." and, "Similarly, the choice of the rate of discount should in this case be independent of considerations of risk" (p. 366).

The fact that the risk of any given individual project considered alone, and suitably apportioned, is negligible has nothing to do with the rate of discount to be used by government. If Congress appropriates too small an amount for the marginal internal rate of return on public funds to equal a risk free discount rate then, as shown elsewhere,<sup>1</sup> it is simply not efficient to use such a rate. The rate of discount is dependent on the supply of government funds and the uses to which they may be put (including return to the private sector), not on a discount rate specified from outside.<sup>2</sup>

It is possible to interpret the Arrow-Lind analysis another way. The risk free discount they advocate may be taken as a round-about way of arguing that government should so change its budget as to make the marginal internal rate of return equal, in

view of projects available, to the risk free discount rate. However, if this is their point, their analysis is in no way supportive. We must compare larger versus smaller budgets with taxpayer risks, and it's not clear that a larger budget, pooling constant, will not increase such risk. Indeed, in terms both of mean-variance analysis and the Arrow-Lind analysis itself, this is just what we should expect. In the case of mean-variance analysis the taxpayers' position reduces to what Friedrich Lutz and Vera Lutz called "the entrepreneur's own capital," and here variance clearly increases with size of commitment, pp. 193 ff. For Arrow-Lind we have the same result since expanding size of commitment is commensurate with reducing the number of persons over whom the risk is distributed.

Thus two conclusions emerge. One, their analysis tells us nothing about what discount rate is appropriate in the framework of budgetary constraint. Two, their analysis is inadequate to support the judgment that the budgetary constraint *should* be consistent with a risk free discount rate.

There remains the implicit bias for reimbursing government projects in the Arrow-Lind discussion. Benefits accruing to individuals are to be discounted at the risk rate applicable to the individual while publicly borne costs will be discounted at the risk free government rate, with, as they note, the implication that "... for the typical case where costs are borne publicly and benefits accrue privately, this procedure will qualify fewer projects than the procedure of using a higher rate to discount both benefits and costs" (p. 378). From this it would follow that reimbursing projects, i.e., projects involving sales, would be favored, since the risk free discount rate would apply to benefits. This result is comparable to Otto Eckstein's case for high capital in-

\* Department of economics, Central Michigan University.

<sup>1</sup> See Nichols (1964).

<sup>2</sup> Thus the private sector rate of return would be the *minimum* cut-off, as in Nichols (1969). In Arrow-Lind presumably the risk free discount rate would play that role.

tensity projects (pp. 62-65 and 70-78) and E. J. Mishan's "reinvestment" of consumption benefits at a lower earnings rate than recaptured benefits, though it has a different justification.

Budget-augmenting projects are favored by the Arrow-Lind criterion because they involve less risk. But there are just as good reasons for particularly opposing budget-augmenting projects. They lay the basis for expanding government activities by indirect means; they bring revenues but escape the political scrutiny applied to taxes. They tend to take government into activities that are already provided by business and thus shift government away from what it's presumably best qualified to cope with, externalities. Granting the Arrow-Lind risk analysis (for individual projects), it thus does not follow that we should on balance shift in the direction of favoring reimbursing projects.

The present criticism of Arrow-Lind does not, of course, bear on their refutation of Hirshleifer nor on their analysis of risk in respect to individual projects. Further, it does not suggest that a larger budget implies a higher discount rate than a smaller budget. Thus a larger budget permits a

lower marginal internal rate of return for a given set of projects. All that is implied is that risk is affected by size of budget and hence it is not correct to argue, as do Arrow and Lind, that government should ignore risk. Finally, if risk aversion is assumed, a modern day "principle of increasing risk" is perfectly consistent with Arrow-Lind and can readily be used to complete a simultaneous solution for budget size, project selection and discount rate.

# REFERENCES

- K. Arrow and R. Lind, "Uncertainty and the Evaluation of Public Investment Decisions," *Amer. Econ. Rev.*, June 1970, 60, 364-78.
- O. Eckstein, *Water-Resource Development*, Cambridge, Mass. 1958.
- F. Lutz and V. Lutz, *The Theory of Investment of the Firm*, Princeton 1951.
- E. J. Mishan, "Criteria for Public Investment: Some Simplifying Suggestions," *J. Polit. Econ.*, Apr. 1957, 75, 1939-46.
- A. Nichols, "On the Social Rate of Discount: Comment," *Amer. Econ. Rev.*, Dec. 1969, 59, 909-11.
- , "The Opportunity Cost of Public Investment: Comment," *Quart. J. Econ.*, Aug. 1964, 78, 499-505.

# Uncertainty and the Evaluation of Public Investment Decisions: Comment

By DONALD WELLINGTON\*

Kenneth Arrow and Robert Lind have argued that public investment should be discounted at a lower rate than private investment. The reason is that public investments are financed by taxpayers who make up a much larger group than any group of private investors. Consequently, any risk associated with a public investment is spread over such a large group that the possible loss in income is so small for each taxpayer that he will ignore it. The result is that a public investment can be discounted at a lower rate even though the actual risk of loss for the entire group is the same as in a private investment. If the taxpayers are completely indifferent to the possible loss of income, the appropriate rate of discount for the public investment is the same as that for riskless investments.

The principle has, however, very little practical usefulness because it cannot be extended beyond the case where a public investment is evaluated in isolation. In order for it to be generally used, the taxpayer must be indifferent over the use and fate of a huge slice of his income. Nor is the principle of any use to working politicians. Neither the political leaders nor their subordinates can be indifferent over the varying riskiness of their actions and ventures. Their political fates partly ride on their record of success in their ventures and Arrow and Lind do not have to dredge laboriously through the ancient records of mankind to find instances of political eclipse by leaders whose adventures yielded little advantage to their communities.

The usefulness of their principle is not the only thing that can be questioned. Their principle is only one of the implications of their argument. The major premise of the argument harbors another question. Why is the number of participants larger in public

investments? Although many reasons can be conceived, there is only one significant reason in the real world. It is that the government can use force.

Politics is concerned with resolving conflicts between humans and any political decision is virtually certain to give rise to discontent on the part of some people. These injured people do not want to abide by the decision. It may strike at the crux of their well-being, as occurred to the German Jews whose taxes helped finance the construction of Hitler's extermination chambers. Notwithstanding the injury that can be wrought by a political decision, the injured people usually abide by it. They conform because the armed might of the state stands ready to compel them.

In the case of private and public investments that are comparable in all respects except the degree of participation in their financing, the only important reason for the higher degree of participation in the public investment is that compulsion is being relied upon to secure the participation. Consequently, it cannot be presumed that the private and public investments give the same return. The safest presumption is that if the public investment brings in additional participants, the return on the investment is estimated or valued, by the participants on the average, as lower.

Although these remarks have not constituted a discussion that is primarily economic in nature, they are germane to any evaluation of public ventures, including public investments. They have merely stressed some elemental realities of politics which are, perhaps, matters of small moment to either Arrow and Lind or the bulk of their audience.

## REFERENCE

- K. J. Arrow and R. C. Lind, "Uncertainty and the Evaluation of Public Investment Decisions," *Amer. Econ. Rev.*, June 1970, 60, 364-78.

\* Assistant professor, University of Cincinnati. I wish to thank my colleagues, Joong-Koon Lee and Lloyd Valentine, for helpful comments.

# Uncertainty and the Evaluation of Public Investment Decisions: Reply

By KENNETH J. ARROW AND ROBERT C. LIND\*

In three of the four comments on our paper (Roland McKean and John Moore, E. J. Mishan, and Alan Nichols) the authors extract and dispute the implication that public investment is always better than private because of the government's superior ability at bearing risks. Actually, it was not our intention to discuss the boundaries of investment sectors as between the public and private spheres. Rather, we assumed (and should have stated explicitly) that the division of commodity services between the two spheres has already been decided on other grounds, such as the existence of externalities or other market failures. Public investment is simply investment needed to produce public goods. The problem we were addressing ourself to was that of the appropriate level of this public investment, consideration being given to the displacement of alternative private investment. (We note that Mishan has recognized this possible interpretation of our paper.)

To be sure, the economic factors in allocating an activity to the public or the private sphere are often equivocal. Arguments for socialization based on the failure or inefficiency of the market in the presence of externalities or increasing returns must be balanced against the countering arguments that public production decisions are made with less information about the benefits of individuals (as McKean and Moore and Wellington have reminded us) and frequently with less incentives to efficient decision making. Our argument does, it is true, add one more factor to the case for putting an activity in the public sphere; but it is only one factor among many and does not lead to McKean and Moore's all-or-none conclusion, "if [our] proposition is correct, then *all* investments should be undertaken

by government or *all* carried out privately" (emphasis in the original).

We now turn to other comments. McKean and Moore seek to push our final point, that uncertain benefits to private individuals be valued at less than their expected value, to a *reductio ad absurdum* by suggesting as a parallel that uncertain privately borne costs also be discounted. The point is clever and amusing but unfortunately incorrect. The correct parallel to our discussion of benefits is that uncertain costs should be equated to a certain cost *greater* than their expected value. Indeed, this is stated explicitly in a sentence evidently overlooked by McKean and Moore: "On the other hand, if net benefits to an individual are negative, this requires discounting expected returns at a rate lower than the certainty rate" (Arrow and Lind, p. 378).

Mishan's note gives an excellent restatement of some of the issues. We think though that a proper use of the Kaldor-Hicks criterion can dissipate the mild fog of paradox that Mishan attaches to it in his example. The possibility of social gain through increased public investment is, as always, tied to an imperfection of the private market. In this case, the market that fails is that for risk-bearing. A transfer of resources to the private sector would yield a Kaldor-Hicks improvement only if it included a transfer of risk-bearing also. But the crucial point in our argument is that only the government can bear risks adequately, through its taxing power.

Mishan's Section III is a little surprising in view of his widely known depreciation of output as a welfare measure. If greater output were obtained at the expense of greater noise, Mishan would not argue that the noise should be disregarded in the evaluation of welfare. Individuals manifest in many ways a dislike for uncertainty; why should not risk abatement be given value in eco-

\* Professor of economics, Harvard University and director, Institute for Public Policy Analysis, Stanford University, respectively.

conomic calculations along with noise abatement?

There is really little to disagree with in Nichols' comments, once the misunderstanding above has been noted. He suggests the use of some internal rate of return criterion if the government budget is limited, but quickly recognizes an alternative (and correct) interpretation of our argument, that it is the size of the public investment budget which is at stake.

We do not quite follow his arguments under the second interpretation. If we understand him correctly, he says that the larger budgets which would follow from our recommendations would involve greater risks, and we do not therefore know *a priori* that the new situation is preferable. But after all it is the burden of our argument that for public investment there are approximately no costs associated with uncertainty and therefore only the mean should matter. Indeed, our proposals imply a shrinking of the risky private investment budget and therefore a reduction in the amount of risk borne by some individuals.

Donald Wellington's polemics make for good reading, though their relevance to careful economic analysis is not entirely

clear. We do not suppose that "the taxpayer is supposed to be indifferent over the use and fate of a huge slice of his income." Public investment policy in a democratic society is supposed to reflect the desire of the typical citizen to achieve a social good yielded by an investment project. We are simply pointing out that since the government can spread the risks better than any private firm, the government should push the margin of its investment further than a private firm can or should. Admittedly, the government's ability to spread risks rests on its powers of compulsion, but this rests in turn on consent. Wellington's rhetoric, if taken at face value, implies the abolition of the state, a proposition that has much to commend it; but it does seem rather a big issue to tie on to the riskiness of public investment.

We are glad to note that Wellington extends to political leaders our principle that business managers may be unduly risk averse (i.e., unduly from the point of view of the stockholders). But we are a little surprised; it has been more usual to hold that political leaders, whose term of office may expire long before the fruits of their investment projects are fully known, have too little rather than too much risk aversion.

# Optimal Taxes and Pricing: Comment

By YEW-KWANG NG\*

In a recent issue of this *Review*, there are three articles on the problem of constrained optimum<sup>1</sup> in pricing or taxation (William Baumol and David Bradford, Avinash Dixit, and Abba Lerner). There is an inconsistency between the paper by Baumol and Bradford (hereafter referred to as BB) and that by Dixit. The latter has shown that, if all goods are taxable, a proportional tax structure (i.e., prices proportional to marginal costs) is optimal. This must also be the conclusion of Lerner if his argument is extended to the case where there is no untaxable sector. However, BB contend that this is not true, arguing that government revenue must be zero with a universal proportional tax. The purpose of this note is to explain this inconsistency.

The argument of BB is reproduced as below:

In formal terms, let  $P_i$  be the price of commodity  $i$  ( $i=1, \dots, n$ ) facing the consumer, . . . and let  $P_L$  be the price of labor. Following the sign convention that positive numbers represent net purchases of desired goods, negative numbers, net sales, let  $a_1, \dots, a_n, a_L$  be the vector of net transactions carried out by a consumer. Our no-lump sum tax assumption can be expressed as the requirement that  $P_1 a_1 + P_2 a_2 + \dots + P_n a_n + P_L a_L = 0$ . . . Obviously, here it makes absolutely no difference to the consumer whether he faces prices  $P_1, \dots, P_n, P_L$  or prices ten times as high; only relative prices count. If the government were to collect the difference between these two price vectors ( $P$  and  $10P$ ) as a tax, it would collect precisely  $9(\sum_i P_i a_i + P_L a_L)$ , which is zero. A tax vector which is a scalar multiple of the price

vector facing consumers will, under our zero-transfer assumption, *always* yield zero revenue. [p. 276]

The above argument is based on the implicit assumption that, for universal proportional tax or marginal cost pricing, the price of labor facing the consumer must be made exactly the same multiple of its marginal cost as all other commodities. Thus the government has to subsidize the consumer (the suppliers of labor) for the exact degree in which it taxes other commodities. This will leave the government with zero net revenue.

However, it should be noted that the required equal proportionality applies only to final goods and not to intermediate goods. Thus, if coal is taxed  $k$  percent no matter if it is used for heating (final good) or for manufacturing steel (intermediate good), and if steel is also taxed the same  $k$  percent, the price of steel will exceed its real marginal cost by a greater degree than does the price of coal. Thus to achieve equal proportionality, the taxes should only be placed on final goods but not on intermediate goods. Prices universally proportional to marginal costs refers only to final goods.

Now labor is used to produce other goods and hence is an intermediate good. Instead of labor, it is leisure which is a final good and must be taxed to achieve universal proportionality.<sup>2</sup> Where leisure is not taxed, the price of leisure to the consumer is the earning forgone. Where other final goods are taxed, leisure must also be taxed to achieve universal proportionality. This will raise the price of leisure which will then be the earning forgone plus the tax. The proportion of the price of leisure to its marginal cost can then be made equal to that proportion for all other final goods.

With the above proposed structure of pro-

\* Senior lecturer in economics, The University of New England, Armidale, Australia. Herbert Mohring's paper, which contains an appendix commenting on the same problem in a somewhat different way, was available to me only after this note had been written.

<sup>1</sup> The optimum is constrained by the requirements that certain revenue be raised by the government and that lump sum tax is not feasible.

<sup>2</sup> I disregard of course the practical difficulty of taxing leisure. As I am dealing with the case where there is no untaxable sector, this question is really irrelevant.

portional taxes, not only can optimum be achieved, but the government can also collect any amount of revenue it needs within the confines of productive capacity and political feasibility. It taxes all final goods including leisure by the same proportion, but it does not have to subsidize anything. It must, therefore, be able to collect a positive revenue. The prices for all final goods facing the consumer are raised by the same proportion but the price of labor is not. He has therefore to cut down his overall consumption, thus releasing resources to the public sector.

It seems therefore that while Dixit's conclusion of a proportional tax structure is correct, he fails to point out that it is only applicable to final goods. This is implicitly contained in Dixit's paper as all commodities enter directly into the utility function of the

consumer. On the other hand, BB's conclusion that taxes "cannot be proportional to consumer prices if there is to be any tax yield," p. 276, is misleading.

#### REFERENCES

- W. J. Baumol and D. F. Bradford, "Optimal Departures from Marginal Cost Pricing," *Amer. Econ. Rev.*, June 1970, 60, 265-83.
- A. K. Dixit, "On the Optimum Structure of Commodity Taxes," *Amer. Econ. Rev.*, June 1970, 60, 295-301.
- A. P. Lerner, "On Optimal Taxes with an Un-taxable Sector," *Amer. Econ. Rev.*, June 1970, 60, 284-94.
- H. Mohring, "The Peak Load Problem with Increasing Returns and Pricing Constraints," *Amer. Econ. Rev.*, Sept. 1970, 60, 693-705.

# Optimal Taxes and Pricing: Reply

By DAVID F. BRADFORD AND WILLIAM J. BAUMOL\*

The problem to which Yew-Kwang Ng addresses himself has also troubled others including Herbert Mohring, who wrote about it in an appendix to his paper on peak load pricing in this *Review*. We will show that our difference with Abba Lerner, Mohring, and Ng is a matter of definition which implicitly leads to a different view of the basic problem at issue. The question is what one means by "a tax on all goods"—whether one defines it in such a way as to make it completely inescapable, thereby implicitly making it possible to impose a tax that produces no misallocation of resources. The difference between our results and Dixit's must be resolved differently.

The definitional problem does not reside in the connotation we assign the term "leisure." All of us apparently define it as that portion of the unit time which the individual does not spend working. Rather, the disagreement resides in the definition of "a commodity tax." We define such a tax, p. 276, to fall on a net transaction, e.g., the sale of a commodity. If one prefers, we have no objection to a universal tax being taken to cover do-it-yourself pursuits, i.e., "transactions with nature" which are not actually carried out on markets. However, our definition requires the individual to have *some options* that can affect the size of his tax bill. Put the other way, the universal tax must, on our definition, permit the individual at least sometimes to refrain from taxed economic activities.

Under this definition of the problem, each individual is endowed with certain flows of "goods" (in the simplest model, 24 hours of leisure per day) some of which he sells in order to buy others. What is taxed is every such *change* in his bundle of goods. Unless the individual receives transfer income, the net *value* of this vector of changes must be zero (this is the budget constraint); hence our conclusion, as summarized by Ng.

\* Princeton University.

We have two reasons for preferring this definition. First, it does seem to be a better description of most (though not all) taxes one encounters in reality. For most taxes do strike transactions rather than untraded holdings or flows.

More important, we prefer "our" definition because we believe using the other assumes away the problem with which the discussion is concerned. A budgetary constraint affects the optimal tax structure only because no lump sum taxes are possible. Otherwise the optimal tax *is* the lump sum tax because it produces no distorting incentive effects. But if it is possible to tax *both* leisure sold (labor) *and* leisure retained, the authorities can simply tax the individual's existence at some fixed rate per hour. It is then, clearly, trivial to convert the commodity tax into a lump sum tax. Note that even the assumption that only labor and leisure are taxable permits the problem to be assumed away in this manner.

Lerner, Mohring, and Ng, all with varying degree of explicitness, interpret the "taxability of all goods" to mean that the individual cannot escape taxation by refraining from consumption or from some other combination of activities, and the optimality of a tax on the inescapable activities (the lump sum tax) follows at once. But none of these authors seem to recognize the reason the result follows so simply under their definition.<sup>1</sup>

Avinash Dixit's result has a different basis. Since he and we use the same analytic framework and the same definitions our difference cannot arise from this source. The difference lies rather in his assumption that

<sup>1</sup> Mohring's analysis of his simple example may even appear to lead to a different answer, since he says that the first best solution requires "... that the tax on a leisure hour be set equal to the sum of the taxes on a unit of commodity *X* and on one working hour" (p. 705). However, reflection will convince the reader that the same optimality conditions require the tax on *X* to be zero (i.e., that the multipliers  $\mu$  and  $\lambda$  must be equal).

consumers enjoy a nonzero net transfer income,  $I$  (see expression (1), p. 296). If  $I$  is positive it is possible to collect positive taxes, in principle up to the full amount of  $I$ , using taxes proportional to prices. The question is, whence cometh  $I$ ? If there are no profits in the system (e.g., in a competitive economy with constant returns to scale) then aggregate  $I$  must be zero unless it is provided through taxes. Since these are not permitted to be lump sum we are back with  $I$  equals zero, in which case taxes proportional to prices raise no revenue. If there *are* profits in the system (an interesting case, of course, but one violating the no-net-transfer assumption), taxes proportional to prices *do* amount to a profit tax, a point missed by Dixit (see p. 301), and do permit a first

best solution up to the point where the total tax take exceeds the profitability of the system.

#### REFERENCES

- W. J. Baumol and D. F. Bradford, "Optimal Departures from Marginal Cost Pricing," *Amer. Econ. Rev.*, June 1970, 60, 265-83.
- A. K. Dixit, "On the Optimum Structure of Commodity Taxes," *Amer. Econ. Rev.*, June 1970, 60, 295-301.
- A. P. Lerner, "On Optimal Taxes with an Un-taxable Sector," *Amer. Econ. Rev.*, June 1970, 60, 284-94.
- H. Mohring, "The Peak Load Problem with Increasing Returns and Pricing Constraints," *Amer. Econ. Rev.*, Sept. 1970, 60, 693-705.

# The Role of Money in a Simple Growth Model: Comment

By JON HARKNESS\*

In a recent article in this *Review*, David Levhari and Don Patinkin (L-P) develop an interesting growth model for a simple economy in which money is treated as a producer's good, entering the aggregate production function. As a policy implication, they find that a rise in the rate of monetary expansion has indeterminate effects on equilibrium capital intensity and real cash balances. The purpose of this note is to briefly demonstrate that these effects are, in fact, determinate and positive, given local stability. This result comes directly from the L-P model and requires no further assumptions or respecifications. L-P have simply failed to carry through their analysis.

The L-P model can be briefly summarized. Assume the economy can be represented by a neoclassical growth model with an aggregate production function  $Y = G(K, N, M/P)$  where  $Y$ ,  $K$ ,  $N$ , and  $M/P$  are output, the physical capital stock, the effective labor force, and real cash balances, respectively. Assume all three factors are cooperant and that the production function is linear homogeneous so that it can be written in the intensive form,  $y = g(k, m)$ , where  $y$ ,  $k$ , and  $m$  are  $Y/N$ ,  $K/N$  and  $M/PN$ , respectively. Given that  $g$  is a well-behaved function, there will be positive but declining marginal products. Thus:  $g_k > 0$ ;  $g_m > 0$ ;  $g_{kk} < 0$ ;  $g_{mm} < 0$ ; and  $g_{km} = g_{mk} > 0$ . The effective labor force grows at the exogenous rate  $\dot{N}/N = n$ .

The government expands the nominal money supply at the rate  $\dot{M}/M = u$  and injects money into the economy by way of transfer payments. This is fiat money with zero production and distribution costs. No government purchases are undertaken. If the rate of inflation is  $\dot{P}/P = \pi$ , the growth of real money balances will be  $(\dot{M}/P) = (u - \pi)M/P$ . Real disposable income is found to be the sum of output plus the

growth of the real money supply (ie., transfer payments). Thus,  $Y_d = Y + (u - \pi)M/P$ . A fixed proportion,  $s$ , of disposable income is saved and, of course, real saving must equal the growth of real wealth, which is comprised of the growth of the physical capital stock,  $\dot{K}$ , and the growth of real cash balances,  $(u - \pi)M/P$ . Thus:

$$(1) \quad s[G(K, N, M/P) + (u - \pi)M/P] \\ = \dot{K} + (u - \pi)M/P$$

This is L-P's equation (42), p. 738. It is readily shown that  $\dot{K}/N = \dot{k} + nk$ . Making use of this relationship, divide (1) by  $N$  and rearrange terms to obtain the following expression for the dynamic path of capital intensity.

$$(2) \quad \dot{k} = s \cdot g(k, m) + (s - 1)(u - \pi)m - nk$$

Since  $m = M/PN$ , it is readily shown by L-P that the rate of growth of real cash balances will be:

$$(3) \quad \dot{m}/m = u - \pi - n$$

Finally, firms are the only holders of real balances, the demand for which is derived from the first-order conditions for profit maximization. The marginal product of capital is  $g_k > 0$ . The marginal physical product of real balances is  $g_m > 0$ . However, there are also marginal capital gains on holdings of real balances equal to  $-\pi$ . Hence, L-P argue, the overall marginal product of real cash balances will be  $g_m - \pi$ . Since the price of physical capital and real balances is, by definition, unity, the first-order condition for profit maximization will be:

$$(4) \quad g_k(k, m) = g_m(k, m) - \pi$$

This is L-P's equation (45), p. 739, which can be solved implicitly for the demand for real balances in terms of  $k$  and  $\pi$ . Equations (2) to (4) completely describe the dynamic

\* Northwestern University.

system, which can be solved for the growth paths of  $k$ ,  $m$ , and  $\pi$ . The steady-state equilibrium is found by setting  $\dot{k}$  and  $\dot{m}$  to zero. Making appropriate substitutions, the steady-state equilibrium reduces to:

$$(5) \quad n(k+m) = s[g(k, m) + nm]$$

$$(6) \quad n = u - g_m(k, m) + g_k(k, m)$$

where:

$$g(k, m) + nm = y_d$$

Equations (5) and (6) can be solved for the steady-state values of  $k$  and  $m$ . If we differentiate the system with respect to  $u$ , we obtain the policy multipliers for an increase in the rate of monetary expansion.

$$(7) \quad dk/du = -[sg_k + (s-1)n]/\Delta$$

$$(8) \quad dm/du = [sg_k - n]/\Delta$$

where

$$(9) \quad \Delta = [sg_k - n][g_{mm} - g_{km}] - [sg_m + (s-1)n][g_{mk} - g_{kk}]$$

Equations (7) to (9) are L-P's equations (53) to (55), p. 740. Both these policy effects are indeterminate, according to L-P, because "...for different—though 'quite' reasonable—values of the variables, the expression  $sg_k - n$  . . . can be either positive or negative" (p. 743). Hence,  $\Delta$  is indeterminate. Moreover, this indeterminacy "is not removed by assuming the system is stable." Appeal to the Correspondence Principle is of no help. These assertions are false.

First, we show that  $sg_k - n$  is unequivocally negative. In equilibrium, per capita disposable income is  $y_d = g(m, k) + nm$ . Moreover, the equilibrium money market, (6), gives us  $n = u - g_m + g_k$  and, with a linear homogeneous production function, factor shares must exhaust output so that  $g(k, m) = w + g_m m + g_k k$ , where  $w$  is the real wage rate. Making use of these two results disposable income can be defined as  $y_d = w + g_k(k+m) + um$ . Finally, from equation (5) we see that, in equilibrium,  $n = sy_d/(k+m)$ . Thus,  $sg_k - n = sg_k - sy_d/(k+m) = [sg_k(k+m) - s(w+um)]$

$$-sg_k(k+m)]/(k+m) = -s(w+um)/(k+m) < 0.$$

Second, by appeal to the Correspondence Principle we show that  $\phi = sg_m + (s-1)n$  must be positive and  $\Delta$  must be negative if the system is locally stable. Linearize the dynamic system, equations (2) to (4) around the steady-state equilibrium using Taylor expansions. In matrix form we obtain the following linear dynamic system:

$$\begin{bmatrix} \dot{k} \\ \dot{m}/m \\ 0 \end{bmatrix} = \begin{bmatrix} sg_k - n & sg_m + (s-1)n & (1-s)m \\ 0 & 0 & -1 \\ g_{mk} - g_{kk} & g_{mm} - g_{km} & -1 \end{bmatrix} \cdot \begin{bmatrix} dk \\ dm \\ d\pi \end{bmatrix}$$

where  $dk$ ,  $dm$  and  $d\pi$  represent the deviations of  $k$ ,  $m$ , and  $\pi$  from their respective steady-state values. Let the coefficient matrix of this system be  $A$ . Stability requires that the real parts of the roots of the equation  $\det[A - \lambda b] = \alpha_2 \lambda^2 - \alpha_1 \lambda + \alpha_0 = 0$  be negative, where  $b$  is the 3x3 matrix

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Letting  $a_{ij}$  be the  $ij$ th element of  $A$ , it is readily confirmed that  $\alpha_2 = a_{33} = -1$ . Since the roots of the above equation will all be negative only if  $\alpha_2$ ,  $-\alpha_1$  and  $\alpha_0$  are all the same sign, this means that a necessary condition for stability is that  $\alpha_0 < 0$ . We will find that:

$$\alpha_0 = [sg_k - n][g_{mm} - g_{km}] - [sg_m + (s-1)n][g_{mk} - g_{kk}] = \Delta$$

Given the signs of the second derivatives of  $g$ , assumed above,  $g_{mm} - g_{km} < 0$  and  $g_{mk} - g_{kk} > 0$ . Since we have already shown that  $sg_k - n < 0$ , the sign of  $\alpha_0 = \Delta$  can be negative if, and only if,  $\phi = sg_m + (s-1)n$  is positive. Thus, assuming the system is stable, we have:  $\Delta < 0$  and  $\phi > 0$ . But the

policy multipliers, (7) and (8), obtained by L-P are

$$dk/du = -\phi/\Delta > 0$$

$$dm/du = (sg_k - n)/\Delta > 0$$

Hence, given stability, both these effects are positive. Of course, the L-P assertion that the system is not unequivocally stable still holds since it can not be determined, a priori, if  $\phi$  will be sufficiently large to guarantee that  $\Delta < 0$ . Moreover, stability also requires that the coefficient  $\alpha_1$  of the characteristic equation be positive, while it is, a priori, indeterminate.<sup>1</sup> However, without

<sup>1</sup> The value of  $\alpha_1$  can be shown to be  $n - sg_k - (g_{mk} - g_{kk}) \cdot (1-s)m + (g_{mm} - g_{km})$ , which is, a priori, indeterminate. For stability it must be positive.

stability, investigation of policy effects by means of comparative dynamics is meaningless since the perturbed system will never generate a new steady-state equilibrium.

In short, with a stable Levhari-Patinkin monetary growth model in which money is a factor of production, a rise in the rate of expansion of the nominal money supply will unequivocally raise the equilibrium level of capital intensity and real cash balances. At least with respect to capital intensity, this is in accord with other "Tobinesque" monetary growth models.

#### REFERENCE

- D. Levhari and D. Patinkin, "The Role of Money in a Simple Growth Model," *Amer. Econ. Rev.*, Sept. 1968, 58, 713-53.

# The Role of Money in a Simple Growth Model: Comment

By R. RAMANATHAN\*

David Levhari and Don Patinkin introduced a model of economic growth in which money is treated as a consumer's good. They showed that a unique balanced-growth path exists in which the long-run equilibrium capital-labor ratio could be larger than the equilibrium capital intensity of the Solow-Swan model (see T. W. Swan and Robert Solow). They also demonstrated that if the aggregate price level changes at a constant rate and the demand for real balances with respect to the money rate of interest is inelastic, then the balanced-growth path is also stable. In this paper we examine the long-run properties of the model in the general case when the rate of price change is variable. It is shown without making any additional assumptions that the system is unstable. A possible way in which the model might become stable is examined briefly. Furthermore, it is found that the empirical plausibility of obtaining an equilibrium capital intensity higher than in the Solow-Swan case is open to doubt.

The model consists of the following equations.

- (1)  $Y = F(K, L) = Lf(k)$
- (2)  $k \equiv K/L$
- (3)  $S = s \left[ Y + \frac{M}{P} (\mu + r) \right]$
- (4)  $I = S - \frac{M}{P} (\mu - \pi)$
- (5)  $DK = I - \delta K$
- (6)  $r = f'(k)$
- (7)  $DL/L = n$
- (8)  $M/PK \equiv m = \lambda(i)f(k)/k$

$$(9) \quad i = r + \pi$$

$$(10) \quad DM/M = \mu$$

$$(11) \quad DP/P = \pi$$

The first equation is the production function relating output  $Y$  to the levels of capital  $K$  and labor  $L$ . It is homogeneous of degree 1 in capital and labor and has the usual neo-classical properties. Per capita output  $Y/L$  can therefore be expressed as a function of the capital-labor ratio  $k$ . Labor supply grows at the rate  $n$ . Saving  $S$  is a constant fraction of disposable income. The assumption of constancy of the saving rate is later relaxed. The nominal money supply  $M$  grows at the rate  $\mu$ , and  $P$  is the price level which changes at the rate  $\pi$ . The real rate of return to capital ( $r$ ) is the marginal product of capital  $f'(k)$ . Disposable income is obtained as the output  $Y$  plus the real value of government transfer payments ( $\mu M/P$ ) less the loss in real value of existing cash balances due to inflation ( $\pi M/P$ ) plus the opportunity cost of holding money which is  $(r + \pi)(M/P)$ . The last term needs some explanation.

Levhari and Patinkin have argued that any rationale for holding money ought to interpret money balances either as a consumer's good or as a producer's good. We will consider only the first case here. If money is treated as a consumer's good then it means that people derive utility from money holdings. The interpretation is that money provides utility by offering protection against uncertainties. Under this interpretation the cost of holding money balances should be included in disposable income just as the cost of holding a commodity is included in disposable income. The opportunity cost of holding money is the money rate of interest ( $i$ ) which is the real return to capital ( $r$ ) plus the loss through inflation at the rate  $\pi$ .

\* Assistant professor of economics, University of California, San Diego. I am thankful to John Conlisk for valuable comments on an earlier draft.

The disposable income is therefore  $Y + (M/P)(\mu + r)$  from which the level of saving is obtained as in equation (3). Not all of this saving can go into gross investment ( $I$ ). The rate of change of real balances  $(\mu - \pi)M/P$  cannot be held in the form of physical assets. Therefore gross investment is saving less this amount. Capital accumulation ( $DK$ ) is net investment as given by equation (5).<sup>1</sup>

Levhari and Patinkin assume the following equation for the demand for money ( $M_d$ ).

$$(12) \quad \frac{M_d}{P} = \lambda Y$$

where

$$\lambda = \lambda(i), \quad \lambda'(i) < 0$$

The ratio of real balances to aggregate output depends on the opportunity cost of holding money. If  $r$  or  $\pi$  increases this cost increases and therefore the demand for money decreases. Price level adjusts instantaneously to equate the demand and supply for real balances. This leads to equation (8) in which both sides have been divided by  $K$ .

The system has eleven simultaneous equations in the eleven endogenous variables  $Y, K, L, S, I, M, P, k, r, i$ , and  $\pi$ . The parameters of the system are  $s, n, \delta$ , and  $\mu$ . Note that  $\pi$  cannot also be treated as a parameter. We would then have an overdetermined system. Levhari and Patinkin have examined the case in which  $\pi$  is treated as a parameter and  $\mu$  is determined endogenously.<sup>2</sup> Let  $\sigma$  be the gross investment-income ratio  $I/Y$ . It is easily shown that  $\sigma$  is a function of  $k$  and  $\pi$  and has the following expression.

$$(13) \quad \sigma = s + \lambda[si - (1 - s)(\mu - \pi)]$$

The rate of growth of capital stock is given by

$$DK/K = \sigma a(k) - \delta$$

where  $a(k) = f(k)/k$  the average product of capital. Under the neoclassical assumptions

<sup>1</sup> Throughout this paper the time derivative of a variable  $X$  is denoted by  $DX$ .

<sup>2</sup> The question of neutrality of money with respect to changes in  $\mu$  arises only when  $\mu$  is treated as a parameter.

the average product curve is downward sloping. Thus  $a'(k) < 0$ . Since  $Dk/k = DK/K - n$  we have the following basic differential equation in  $k$ .

$$(14) \quad Dk/k \equiv \phi(k, \pi) = \sigma a(k) - (n + \delta)$$

Logarithmic differentiation of (8) with respect to  $t$  gives

$$\begin{aligned} \mu - \pi - DK/K &= \frac{k}{m} \frac{\partial m}{\partial k} \frac{Dk}{k} + \frac{1}{m} \frac{\partial m}{\partial \pi} D\pi \\ (15) \quad \mu - \pi - (Dk/k + n) &= (\epsilon - 1) Dk/k + \frac{\lambda'}{\lambda} D\pi \end{aligned}$$

where  $\epsilon$  is the partial elasticity of demand for real balances with respect to the capital stock.

$$\begin{aligned} \epsilon &= \frac{K}{(M/P)} \frac{\partial (M/P)}{\partial K} = \frac{1}{m} \frac{\partial}{\partial K} (Km) \\ (16) \quad &= 1 + \frac{k}{m} \frac{\partial m}{\partial k} \end{aligned}$$

In deriving (15) we have made use of the fact that  $(1/m)(\partial m/\partial \pi) = \lambda'/\lambda$ . The following differential equation in  $\pi$  is readily obtained.

$$\begin{aligned} D\pi &\equiv B(k, \pi) \\ (17) \quad &= \frac{-\lambda}{\lambda'} [\pi - \mu + n + \epsilon Dk/k] \end{aligned}$$

In the steady state  $Dk = D\pi = 0$ . From (17) we have the well-known result  $\pi^* = \mu - n$ . Equations (14) and (17) are the basic differential equations of the system and may be rewritten as follows:

$$(18) \quad Dk = A(k, \pi) \quad \text{and} \quad D\pi = B(k, \pi)$$

Let  $(\pi^*, k^*)$  be the steady-state solution given by the relations  $A(k, \pi) = 0$  and  $B(k, \pi) = 0$ . Levhari and Patinkin have shown that such a solution exists, is unique, and that it is possible for the steady state  $k^*$  to be higher than the  $k^*$  in the Solow-Swan model.

To analyze the stability of the system in the neighborhood of the equilibrium, set

$k = k^* + \Delta k$  and  $\pi = \pi^* + \Delta \pi$ . The increments  $\Delta k$  and  $\Delta \pi$  are assumed to be small, that is, their higher powers can be neglected. Substituting these in (18) we get

$$\begin{aligned} D(\Delta k) &= A(k^* + \Delta k, \pi^* + \Delta \pi) \\ &= A(k^*, \pi^*) + b_{11}\Delta k + b_{12}\Delta \pi \end{aligned}$$

to the first-order of approximation. The coefficients  $b_{11}$  and  $b_{12}$  are the partial derivatives of  $A(k, \pi)$  with respect to  $k$  and  $\pi$ , respectively, evaluated at  $(\pi^*, k^*)$ . Since  $A(k^*, \pi^*) = 0$  this can be written as

$$(19) \quad D(\Delta k) = b_{11}\Delta k + b_{12}\Delta \pi$$

Similarly,

$$(20) \quad D(\Delta \pi) = b_{21}\Delta k + b_{22}\Delta \pi$$

where  $b_{21}$  and  $b_{22}$  are similar partial derivatives of  $B(k, \pi)$ . It can be shown (see Richard Bellman, pp. 244-45) that the necessary and sufficient conditions for the stability of the linear system (19) and (20) are

$$b_{11} + b_{22} < 0$$

and

$$b_{11}b_{22} - b_{12}b_{21} > 0$$

Since  $A(k, \pi) = k\phi(k, \pi)$  and  $\phi(k^*, \pi^*) = 0$  we have

$$\begin{aligned} (21) \quad b_{11} &= k^* \left( \frac{\partial \phi}{\partial k} \right)^* = k^* [a' + a\sigma_k] \\ &= k^* a \left[ \frac{\sigma^2}{n + \delta} a' + \sigma_k \right] \end{aligned}$$

$$(22) \quad b_{12} = k^* \left( \frac{\partial \phi}{\partial \pi} \right)^* = k^* a\sigma_\pi$$

$\sigma_k$  and  $\sigma_\pi$  are the partial derivatives of  $\sigma$  with respect to  $k$  and  $\pi$ , respectively, evaluated at the equilibrium. Similarly, we have from (17)

$$(23) \quad b_{21} = B_k^* = -\frac{\epsilon\lambda}{k\lambda'} b_{11}$$

$$(24) \quad b_{22} = B_\pi^* = -\frac{\lambda}{\lambda'} (1 + b_{12}\epsilon/k)$$

Therefore from (23) and (24),

$$\frac{b_{22}}{b_{21}} - \frac{b_{12}}{b_{11}} = \frac{k}{\epsilon b_{11}} = \frac{-\lambda}{\lambda'} \frac{1}{b_{21}}$$

It follows that  $b_{11}b_{22} - b_{21}b_{12} = (-\lambda/\lambda')b_{11}$ . Since  $\lambda' < 0$ , the second condition for stability will be satisfied if, and only if,  $b_{11} > 0$ . Differentiating (13) partially with respect to  $k$  and then setting  $\mu - \pi = n$  we get

$$\sigma_k = sf''(\lambda + i\lambda') - (1 - s)n\lambda'f''$$

Let  $\eta$  be the partial elasticity of demand for real balances with respect to the money rate of interest  $i$ .

$$\eta = \frac{-i}{(M/P)} \frac{\partial (M/P)}{\partial i} = \frac{-i\lambda'}{\lambda}$$

Using this we can rewrite  $\sigma_k$  as follows

$$\sigma_k = sf''\lambda(1 - \eta) - (1 - s)n\lambda'f''$$

Similarly,  $\sigma_\pi = s\lambda(1 - \eta) + (1 - s)(\lambda - n\lambda')$ . Levhari and Patinkin assume that the demand is inelastic ( $\eta < 1$ ) and cite a number of empirical studies justifying the assumption. Making this assumption and noting that  $\lambda' < 0$  and  $f'' < 0$  it follows that  $\sigma_k < 0$  and  $\sigma_\pi > 0$ . Also  $a' < 0$  and therefore from (21) and (22)  $b_{11} < 0$  and  $b_{12} > 0$ . This implies that the second stability condition is violated. The system is therefore unstable.<sup>3</sup>

This result can also be presented with the help of a *Phase Diagram*.<sup>4</sup> We just showed that  $b_{11} < 0$  and  $b_{12} > 0$  under the assumption that the demand for real balances with respect to the money rate of interest is inelastic. It follows from this that  $b_{21} < 0$ , and  $b_{22} > 0$ . The slope of the curve representing the relation  $A(k, \pi) = 0$  at the equilibrium point is  $dk/d\pi = -b_{12}/b_{11}$ . This slope is positive because  $b_{11} < 0$  and  $b_{12} > 0$ . Therefore the graph of the  $Dk = 0$  equation is upward sloping. Similarly the slope of the  $D\pi = 0$  equation is  $-b_{22}/b_{21}$  which is also positive.

<sup>3</sup> Because the determinant of  $((a_{ij}))$  is negative, the corresponding characteristic roots will have opposite signs. The solution is therefore a *saddle point*.

<sup>4</sup> Miguel Sidrauski has carried out a similar analysis for a model with price expectations. But he did not treat money as a consumer's good and his assumption regarding the demand for money is more restrictive.

Moreover,  $-b_{22}/b_{21} > -b_{12}b_{11}$ . Therefore, the  $D\pi=0$  curve has a higher slope than the  $Dk=0$  curve. These results are presented in Figure 1.

Since  $b_{11} < 0$ , at all points above the  $Dk=0$  curve  $Dk$  is negative, that is,  $k$  will tend to decrease.<sup>5</sup> Below the curve the opposite is true. Since  $b_{21} < 0$ , above the  $D\pi=0$  curve  $D\pi$  is negative and  $\pi$  will tend to decrease. The directional arrows in the phase diagram indicate these results. It is evident that the system is generally unstable. The only way the steady state can be attained is when the initial position is somewhere on the dotted line. The steady-state solution is thus a *saddle point*.

So far we have assumed that the saving ratio is constant. Levhari and Patinkin assume that  $s=s(r, -\pi)$  with  $s_1 > 0$  and  $s_2 > 0$  where the partial derivatives are taken with respect to the first and second arguments, respectively. It follows that  $s_k < 0$  and  $s_\pi < 0$ . The only change is in the expressions for  $b_{ij}$ 's;  $b_{11}$  and  $b_{12}$  will have additional terms which are all negative. Therefore  $b_{11}$  is still negative and the second condition for stability will be violated. Thus even if the saving function is the more general one the stability condition is not satisfied.

Is there any way the stability conditions may be satisfied? To analyze this, first note that for the second stability condition to be satisfied  $b_{11}$  must be positive. This means that  $b_{22}$  must be negative because otherwise the first stability condition will be violated.  $b_{22}$  cannot be negative unless  $b_{12}$  is negative.  $b_{11}$  is the effect of an increase in  $k$  on the rate of capital accumulation in the neighborhood of the equilibrium. Similarly  $b_{12}$  is the effect of an increase in  $\pi$  on the rate of capital accumulation again near the equilibrium.

Suppose the overall saving ratio is given as  $s=s(r, -\pi, y)$  where  $y=Y/L$  is per capita output. It is assumed that  $s_1 > 0$ ,  $s_2 > 0$ , and  $s_3 > 0$ . Therefore  $s_\pi < 0$  but the result of an increase in  $k$  is ambiguous. It is reasonable to assume and empirically justifiable that

<sup>5</sup> Even though the signs of  $b_{21}$  and  $b_{11}$  have been obtained in the neighborhood of the equilibrium, because the functions  $A$  and  $B$  are continuous, the same signs hold at points not near the equilibrium.

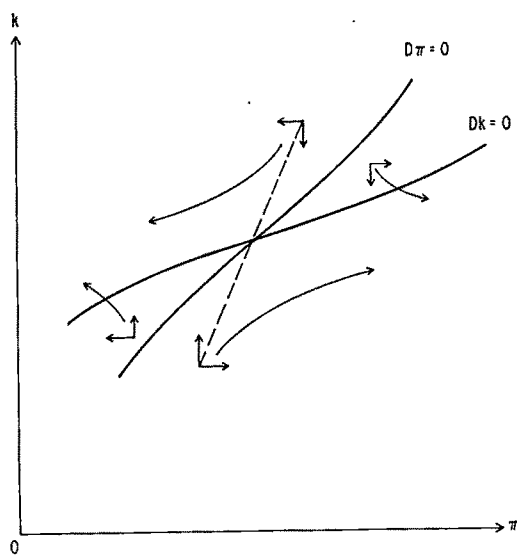


FIGURE 1

income will exert a stronger influence on  $s$  than the real rate of interest. Therefore  $s_k > 0$ . Under this assumption  $b_{11}$  will have additional terms all of which will be positive. If the income effect is sufficiently strong (for a given  $\pi$ )  $b_{11}$  may become positive which implies that the second stability condition may be satisfied.

Under the assumption of a constant saving rate and inelastic demand for money with respect to  $i$ , we showed that  $b_{12} > 0$ . But stability is impossible unless  $b_{12}$  is negative. This is possible only if the extra terms in  $b_{12}$  which arise because of a variable saving ratio are negative and large enough to make  $b_{12}$  not only negative but also make  $b_{22}$  sufficiently negative so that  $b_{11} + b_{22} < 0$ . Since  $s_\pi < 0$  the extra terms are clearly negative. An increase in  $\pi$  (with a fixed  $k$ ) means the real value of existing balances and hence disposable income will decrease. Since output per unit of capital is fixed, consumption will decrease. Saving will therefore increase. On the other hand, an increase in  $\pi$  would reduce the rate of saving and hence reduce the level of saving. The net effect is ambiguous. But stability is impossible unless an increase in  $\pi$  (for a fixed  $k$ ) reduces the saving rate sufficiently enough to make not only the overall rate of capital accumulation (as indicated by  $b_{12}$ )

TABLE 1—MINIMUM MONEY RATE OF INTEREST  
FOR ALTERNATIVE VALUES OF  $n$  AND  $s$ 

$n$	$s$	min $i^*$
.02	.10	.18
(.038)	(.10)	(.34)
.04	.10	.36
.06	.10	.54
.02	.15	.11
.04	.15	.23
(.046)	(.15)	(.26)
.06	.15	.34
.02	.20	.08
.04	.20	.16
(.054)	.20	(.22)
.06	.20	.24
.04	.25	.12
.06	.25	.18
(.062)	(.25)	(.19)

negative but to make  $b_{11} + b_{22} < 0$ . My conjecture is that the effect of  $\pi$  on the saving rate is not much and therefore it will be awfully hard to satisfy the stability conditions.

Before concluding we briefly examine the feasibility of obtaining an equilibrium capital intensity higher than in the Solow-Swan model. In equation (13) set  $i = r + \pi^*$  and  $\pi^* = \mu - n$ . Then

$$\sigma = s + \lambda[s(\mu + r) - n]$$

A larger  $k^*$  than in the Solow-Swan economy will be obtained if, and only if,  $\mu + r > n/s$ . Setting  $\mu = \pi^* + n$  and rearranging terms the condition for a higher equilibrium capital intensity is  $i^* = \pi^* + r^* > n(1-s)/s$ . The feasibility of this condition obviously depends on the values of  $n$ ,  $s$ ,  $\pi^*$ , and  $r^*$ . The gross saving ratio for most countries will probably not exceed 25 percent. Typically it is well below that. The natural rate  $n$  is the rate of growth of labor force plus the rate of growth of technical progress and is the equilibrium rate of growth of aggregate output. Table 1 gives the *minimum* money rate of

interest required for alternative values of  $n$  and  $s$ .

It is found from the table that the minimum rate of interest to achieve an equilibrium capital intensity higher than in the Solow-Swan case is generally high except when the saving rate is large and  $n$  is relatively small. But John Conlisk has shown using cross-country data that the higher the gross saving rate the higher the growth rate.<sup>6</sup> If we assume that labor force grows at an annual rate of 2.5 percent and estimate  $n$  for  $s = .10, .15, .20$ , and  $.25$  we obtain the values presented in parenthesis in Table 1. It is evident that the lower bound for the money rate of interest is unrealistically large. The empirical plausibility of a larger steady state value in the Levhari-Patinkin model is therefore open to doubt.

#### REFERENCES

- R. Bellman, *Introduction to Matrix Analysis*, New York 1960.
- J. Conlisk, "A Modified Neoclassical Growth Model with Endogenous Technical Change," *Southern Econ. J.*, Oct. 1967, 34, 199-208.
- D. Levhari and D. Patinkin, "The Role of Money in a Simple Growth Model," *Amer. Econ. Rev.*, Sept. 1968, 58, 713-53.
- M. Sidrauski, "Inflation and Economic Growth," *J. Polit. Econ.*, Dec. 1967, 75, 796-810.
- R. Solow, "A Contribution to the Theory of Economic Growth," *Quart. J. Econ.*, Feb. 1956, 75, 65-94.
- T. W. Swan, "Economic Growth and Capital Accumulation," *Econ. Rec.*, Nov. 1956, 32, 334-61.
- J. Tobin, "Money and Economic Growth," *Econometrica*, Oct. 1965, 33, 671-84.

<sup>6</sup> Conlisk has estimated the relation  $n = .005 + .155s + .700g$  where  $n$  is the rate of growth of aggregate output,  $s$  is the gross saving ratio, and  $g$  is the rate of growth of population.

# The Role of Money in a Simple Growth Model: Reply

By DAVID LEVHARI AND DON PATINKIN\*

While accepting the technical analysis of R. Ramanathan and Jon Harkness, we would like to point out some inadequacies of interpretation on their part.

Ramanathan's main point is that our model with money in the utility function is unstable. Essentially, the instability of models of this kind was already pointed out by Miguel Sidrauski. The latter showed how stability could be achieved by assuming a price-expectation function with certain properties. (We referred to this very briefly on pp. 731-32 of our original article.) Ramanathan does not refer to this assumption, but instead shows how stability can sometimes be achieved by assuming that the savings *ratio* depends also on the *absolute level* of income—without giving any rationale for this. Indeed, this assumption runs counter to the usual tendency to assume that—in the long run—this ratio remains constant under changes in income.

The discussion at the end of Ramanathan's note is also not satisfactory. For reasons which we indicated on page 717 of our paper, there is not much meaning in comparing a barter economy with a money one. The really meaningful question is the effect on the capital-labor ratio,  $k$ , of a change in the rate of monetary expansion within a given money economy.

Harkness's note deals with the money-in-the-production-function model. There are several points here which deserve further study. Thus Harkness shows that this model can—under certain assumptions—be stable. This raises the question as to whether there is some analytical similarity between these assumptions and those used by Sidrauski and Ramanathan to achieve stability in the utility-function model.

Harkness's main point (in addition to his neat demonstration that  $sg_k - n < 0$ ) is that we have not made proper use of the Correspondence Principle in order to determine

the comparative-statics properties of the system. What Harkness does not realize, however, is that his dynamic system differs from ours. For we have made the simplifying assumption that the rate of monetary expansion is continuously adjusted so as to keep the rate of change of prices,  $\pi$ , constant, whereas Harkness assumes that  $\pi$  as well as  $k$  is free to change during the dynamic process.

This also suggests that the reason Harkness is able to obtain information by use of the Correspondence Principle is that his dynamic system is more general than ours. Hence it would seem that the assumption of stability places more restrictions on the properties of the behavior function in his system than in ours. In any event, there is nothing in Harkness's analysis to contradict the conclusion that the application of the Correspondence Principle to our dynamic system does not provide sufficient information to determine its comparative-statics properties.

One of the puzzling aspects of Harkness's results is that stability implies that an increase in the rate of monetary expansion (and hence in the rate of increase of prices) increases the equilibrium value of real balances (i.e.,  $dm/d\pi > 0$ ). (In terms of our Figure 3, p. 741, there is something in the stability conditions which requires  $cc$  to slope upwards and intersect  $mm$  from above, thus leaving only  $T$  as a possible equilibrium position.) This runs counter to one's expectations—according to which  $dm/d\pi < 0$  is a more reasonable conclusion. Maybe the fault is in the model itself. In any event, the economic meaning of Harkness's conclusion on this point needs further clarification.

## REFERENCES

- D. Levhari and D. Patinkin, "The Role of Money in a Simple Growth Model," *Amer. Econ. Rev.*, Sept. 1968, 58, 713-53.  
M. Sidrauski, "Inflation and Economic Growth," *J. Polit. Econ.*, Dec. 1967, 75, 796-810.

\* The Hebrew University, of Jerusalem.

# A Note on Pollution Prices in a General Equilibrium Model

By LARRY E. RUFF\*

Consider a general equilibrium model in which several types of pollution are produced by a large number of productive units or firms, which also produce ordinary economic goods; the total amount of each type of pollution in the environment is, for now, simply the sum of the outputs from the firms. The total pollution levels are "public bads," which, along with ordinary economic goods, enter into the utility functions of every individual, but have no effect on firms; later, we consider the possibility that pollution affects production functions, and that the pollution which causes damage is actually synthesized by chemical processes in the environment, using effluents from firms as raw materials.

## I

Formally, the model is:

$x^k = (x_1^k, x_2^k, \dots, x_N^k) =$  Net output of ordinary goods for firm  $k$ ;

$z^k = (z_1^k, z_2^k, \dots, z_L^k) =$  Net output of pollutants for firm  $k$ ;

$g^k(x^k; z^k) \geq 0$ , Production constraint for firm  $k$ ,  $k = 1, 2, \dots, K$ ;

$Z_l \equiv \sum_{k=1}^K z_l^k =$  Total pollution of type  $l$ ,  $l = 1, 2, \dots, L$ ;

$q^j = (q_1^j, q_2^j, \dots, q_N^j) =$  Net consumption of goods by individual  $j$ ;

$U^j(q^j; Z) =$  Ordinal utility indicator for individual  $j$ ,  $j = 1, 2, \dots, M$ ;

$\sum_j q_i^j \leq \sum_k x_i^k$ , Accounting constraint for good  $i$ ,  $i = 1, 2, \dots, N$ .

By letting  $L$ , the number of pollutants, be large enough, this model can include all producer-consumer externalities, including the ordinary two-sided interactions of the chimney-laundry type; but we are interested

in the more generalized type of pollution, of which Los Angeles smog is the prototype. Later we will want to assume there are many producers of each type of pollution. We also assume there is always some good an individual desires.

In the absence of any controls on pollutant outputs, we assume a competitive general equilibrium is established, in which pollution levels are considered too high.<sup>1</sup> It is desired to reduce these levels, using economic instruments, i.e., taxes and subsidies. As Charles Plott has pointed out, the taxes and subsidies must be applied directly to pollutant emissions, or something directly and inflexibly related to these emissions; taxing any other variable, such as the firm's output or one of several pollution-influencing inputs, leads to inefficiency as improper substitutions are encouraged.<sup>2</sup> Therefore, we assume the taxes or subsidies are applied directly to emissions.

Ronald Coase has demonstrated convincingly that, except for wealth-transfer effects, there is no difference between a system which taxes emissions, and one which subsidizes emissions below some level. In fact, subsidy systems which truly subsidize emission reductions no matter how accomplished, can be thought of as cash subsidies paid to particular individuals, combined with taxes on emissions.<sup>3</sup> Therefore, we assume control

<sup>1</sup> We are concerned here with the question of efficiency of price equilibria, rather than the more difficult question of existence of such equilibria. However, if consumers' preferred-to sets are convex in  $q$  for every  $Z$ , and production sets are convex, it would not be difficult to prove existence for every vector of desired pollution levels. See Kenneth Arrow.

<sup>2</sup> In fact, if "pollution" is an inferior input for the firms, taxing output may lead to higher pollution and higher costs. Plott uses a model in which pollution is a direct function of only a single input, and hence taxing that input is appropriate; in general, many things influence pollution levels.

<sup>3</sup> Unfortunately, most real-life policies which purport to subsidize pollution control actually subsidize some

\* Economics department, University of California, San Diego.

is imposed in the form of a  $L$ -dimensional pollution fee or tax vector,  $\bar{\tau}$ , with the revenues allocated to individuals in some unspecified way. The system readjusts, reaching a new equilibrium, defined by

**DEFINITION I.** An allocation  $\bar{q}, \bar{x}, \bar{z}$  is said to be a competitive general equilibrium allocation, and  $\bar{p}$  is said to be a competitive general equilibrium price vector, relative to the tax vector  $\bar{\tau}$  and the income distribution<sup>4</sup>  $\bar{b}$ , if:

$$(a) \quad \bar{p}'\bar{q}^j = \bar{b}^j, \quad \bar{q}^j \geq 0,$$

and for all  $q^j \geq 0$  such that  $\bar{p}'q^j \leq \bar{b}^j$ ,

$$U^j(q^j; \bar{Z}) \leq U^j(\bar{q}^j; \bar{Z}), \quad j = 1, 2, \dots, M;$$

$$(b) \quad g^k(\bar{x}^k; \bar{z}^k) \geq 0,$$

and for all  $x^k, z^k$  such that  $g^k(x^k; z^k) \geq 0$ ,

$$\bar{p}'x^k - \bar{\tau}'z^k \leq \bar{p}'\bar{x}^k - \bar{\tau}'\bar{z}^k, \\ k = 1, 2, \dots, K;$$

$$(c) \quad \sum_{k=1}^K \bar{x}_i^k \geq \sum_{j=1}^M \bar{q}_i^j, \quad \bar{p}_i \geq 0,$$

and

$$\bar{p}_i \left[ \sum_{k=1}^K \bar{x}_i^k - \sum_{j=1}^M \bar{q}_i^j \right] = 0, \\ i = 1, 2, \dots, N;$$

$$(d) \quad \bar{Z}_l = \sum_{k=1}^K \bar{z}_l^k, \quad l = 1, 2, \dots, L$$

This definition is the standard one for a competitive equilibrium, with the addition of pollution as a public good. Individuals are choosing their consumption vectors to maxi-

mize their utility, subject to their budget constraint, and ignoring the impact of their consumption choices on pollution levels. Firms are maximizing profits, given prices, taxes and their production constraints. Production of any good does not fall short of consumption, and exceeds it only if the price of the good is zero.

Because of the externality operating through pollution levels, there is no presumption that the competitive price equilibrium allocation is Pareto optimal. It may fail to be Pareto optimal for two distinct reasons. The first is that the levels of pollution may be such that everybody would gain from an increase or decrease in these levels, with the corresponding decrease or increase in income levels; for the moment we dodge the question of choosing the optimal pollution levels. The second possible reason the price allocation may fail to be Pareto optimal is that it may not be efficient, in the sense of

**DEFINITION II.** An allocation will be said to be efficient if it is feasible, and if there is no feasible allocation which has the same aggregate pollution levels and yet is Pareto-preferred, i.e., makes somebody better off without harming anyone.

The definition implies that even if the "correct" pollution levels are achieved, the allocation must be efficient in this sense in order to be Pareto optimal.

It is now possible to prove two simple theorems regarding the efficiency of price equilibria, and the interpretation of the pollution tax rates  $\bar{\tau}$ .

**THEOREM I.** A price equilibrium defined by Definition I is efficient in the sense of Definition II.

**PROOF:**

Let  $\bar{q}, \bar{x}, \bar{z}$  be the allocation of Definition I, and let  $\hat{q}, \hat{x}, \hat{z}$  be any feasible allocation with the same pollution levels, i.e.,

$$\sum_k \hat{z}_l^k = \bar{Z}_l = \sum_k \bar{z}_l^k, \quad l = 1, 2, \dots, L$$

From the profit maximization condition, we know

particular activity which may or may not be directly related to emission reduction. Often, one must establish credentials as a polluter in order to qualify for the subsidies, with the result the plans become subsidies for pollution itself. See Allan Kneese and Kneese and B. T. Bower (1968) for discussions of the practical problems such subsidies face.

<sup>4</sup> We assume individual  $j$  has a budget  $\bar{b}^j$ , which is independent of his consumption decisions. This budget may include lump sum transfers; but individuals as a group must be getting subsidies equal to total pollution fees paid by firms, since the government absorbs no resources in this model. Allocation of these tax revenues provides an obvious income redistribution opportunity.

$$\bar{p}'\bar{x}^k - \bar{\tau}'\bar{z}^k \geq \bar{p}'\bar{x}^k - \bar{\tau}'\bar{z}^k, \\ k = 1, 2, \dots, K$$

Or, summing over all firms and changing the order of summation,

$$\sum_i \bar{p}_i \left[ \sum_k \bar{x}_i^k - \sum_k \bar{x}_i^k \right] \\ \geq \sum_l \bar{\tau}_l \left[ \sum_k \bar{z}_l^k - \sum_k \bar{z}_l^k \right]$$

Since the pollution levels are the same in the two allocations, the right-hand side vanishes, and we have

$$(a) \quad \sum_i \bar{p}_i \sum_k \bar{x}_i^k \geq \sum_i \bar{p}_i \sum_k \bar{x}_i^k$$

Now, suppose the  $\bar{q}$ ,  $\bar{x}$ ,  $\bar{z}$  allocation is Pareto-preferred to the  $\bar{q}$ ,  $\bar{x}$ ,  $\bar{z}$  allocation. By the assumption of nonsatiation in consumption, this implies the  $\bar{q}^j$  bundles must cost more (at prices  $\bar{p}$ ) for some and no less for all consumers than the  $\bar{q}^j$  bundles, and hence the total cost of all bundles must be greater, i.e.,

$$(b) \quad \sum_i \bar{p}_i \sum_j \bar{q}_i^j < \sum_i \bar{p}_i \sum_j \bar{q}_i^j$$

But, since all allocations must be feasible, and prices  $\bar{p}$  are nonnegative, we can say

$$\sum_j \bar{q}_i^j \leq \sum_k \bar{x}_i^k,$$

hence

$$\bar{p}_i \sum_j \bar{q}_i^j \leq \bar{p}_i \sum_k \bar{x}_i^k,$$

and

$$(c) \quad \sum_i \bar{p}_i \sum_j \bar{q}_i^j \leq \sum_i \bar{p}_i \sum_k \bar{x}_i^k$$

Combining (b) and (c) above with (c) of Definition I, we finally obtain

$$(d) \quad \sum_i \bar{p}_i \sum_k \bar{x}_i^k = \sum_i \bar{p}_i \sum_j \bar{q}_i^j < \sum_i \bar{p}_i \sum_j \bar{q}_i^j \\ \leq \sum_i \bar{p}_i \sum_k \bar{x}_i^k$$

Since inequalities (d) and (a) are contradictory, it follows that it is impossible to find an allocation which is feasible, has the same pollution levels, and yet is Pareto-preferred to a price-equilibrium allocation.

**THEOREM II.** *In the general equilibrium of Definition I, the tax-price of pollutant  $l$ ,  $\bar{\tau}_l$ , is the marginal cost, in terms of "national income" as ordinarily defined,<sup>5</sup> and at current prices, of reducing pollutant  $l$ . That is, defining national income at current prices,  $\bar{Y}$ , by*

$$\bar{Y} \equiv \sum_k \bar{p}'\bar{x}^k \equiv \sum_j \bar{p}'\bar{q}^j,$$

it is true that

$$(a) \quad \bar{\tau}_l = \frac{\partial \bar{Y}}{\partial Z_l} \bigg|_{p=p} \equiv \sum_j \bar{p}' \frac{\partial \bar{q}^j}{\partial Z_l} \equiv \sum_k \bar{p}' \frac{\partial \bar{x}^k}{\partial Z_l}$$

**PROOF:**

For simplicity, we will assume continuous, differentiable functions for this proof. Then, the profit maximization conditions imply, for all  $k, i, j, m$ ,

$$(b) \quad \frac{1}{\bar{\tau}_j} \frac{\partial g^k}{\partial z_j^k} + \frac{1}{\bar{p}_l} \frac{\partial g^k}{\partial x_i^k} = 0,$$

$$(c) \quad \frac{1}{\bar{p}_i} \frac{\partial g^k}{\partial x_i^k} - \frac{1}{\bar{p}_m} \frac{\partial g^k}{\partial x_m^k} = 0 \\ g^k(\bar{x}^k, \bar{z}^k) = 0$$

In particular;

$$(b') \quad \frac{1}{\bar{\tau}_j} \frac{\partial g^k}{\partial z_j} + \frac{1}{\bar{p}_1} \frac{\partial g^k}{\partial x_1^k} = 0$$

$$(c') \quad \frac{1}{\bar{p}_i} \frac{\partial g^k}{\partial x_i^k} - \frac{1}{\bar{p}_1} \frac{\partial g^k}{\partial x_1^k} = 0$$

Now consider a differential change in the parameters  $\bar{\tau}$ , to  $\bar{\tau} + \Delta\tau$ . Taking a total differential of  $g^k=0$  yields

<sup>5</sup> The qualification "as ordinarily defined" is required to distinguish this  $Y$  from any measure of "welfare."  $Y$ , as defined here and in ordinary GNP accounts, cannot be considered a reliable index of welfare because it ignores changes in the environment. But changes in  $Y$  can be used as estimates of the "economic" cost of improving the environment.

$$\sum_i \Delta x_i^k \frac{\partial g^k}{\partial x_i^k} + \sum_j \Delta z_j^k \frac{\partial g^k}{\partial z_j^k} = 0$$

Using (b') and (c'), this relation can be rewritten

$$\left( \frac{1}{\bar{p}_1} \frac{\partial g^k}{\partial x_1^k} \right) \left[ \sum_i \Delta x_i^k \bar{p}_i - \sum_j \Delta z_j^k \bar{\tau}_j \right] = 0,$$

from which it follows that

$$\sum_j \Delta z_j^k \bar{\tau}_j = \sum_i \Delta x_i^k \bar{p}_i$$

Or, summing over all  $k$ ,

$$\sum_j \bar{\tau}_j \cdot \Delta Z_j = \sum_j \bar{\tau}_j \sum_k \Delta z_j^k = \sum_i \bar{p}_i \sum_k \Delta x_i^k$$

But, if the tax changes  $\Delta \tau_j$  are chosen so that the aggregate levels of all pollutants except  $l$  are unchanged, i.e., if  $\Delta Z_j = 0, j \neq l$ , then

$$\bar{\tau}_l = \sum_i \bar{p}_i \sum_k \frac{\Delta x_i^k}{\Delta Z_l}$$

from which (a) follows in the limit.

These theorems are really no significant generalization of standard theorems regarding efficiency and competitive prices.<sup>6</sup> This is clearly seen if we regard "waste disposal services" as productive factors, with the specified levels of pollution determining the total amount of services available; that a price system allocates the publicly owned scarce resource efficiently, and that the competitive rental price is the marginal value of the capacity, is no surprise.

## II. Some Extensions of the Simplest Case

The model of Theorems I and II has assumed there are several distinct types of

<sup>6</sup> However, it is worth pointing out that Theorem I is not totally trivial, that there are ways of reducing pollution to a specified level  $\bar{Z}$ , allowing competition in the production of ordinary goods, and getting an allocation which is *not* efficient in the sense of Definition II; it is not always true that, even with given pollution levels, "one cannot help those suffering from pollution without hurting the polluters," since the same pollution levels can be produced in many ways.

pollutant, which are produced by firms and which act directly on consumers. Often, however, the primary effluents interact in the environment to produce synthesized pollutants, which then act on individuals. Smog in Los Angeles is an example, with the more obnoxious components being generated photochemically in the atmosphere, with industrial and automotive emissions providing raw materials. A similar situation arises if it is possible to specify "isodamage" curves or surfaces for some of the pollutants, on the basis of medical findings or cost estimates, for example.

Wherever they come from, suppose we can specify a set of functions of the form

$$S_r = \phi_r(Z), \quad r = 1, 2, \dots, R, \quad R < L,$$

where it is the vector  $S$  which enters into individual utility functions, rather than  $Z$ ; that is,  $U^j(q^j; Z)$  is replaced by  $U^j(q^j; S)$ , but otherwise the model is unchanged. Now, when a set of pollution prices  $\bar{\tau}$  is specified and a general equilibrium is achieved with effluent outputs of  $\bar{Z}$ , the levels of synthesized pollutants (using this interpretation, for concreteness) are  $\bar{S} = \phi(\bar{Z})$ . Theorem II tells us there is no Pareto-preferred way to achieve emission levels  $\bar{Z}$ , but says nothing about better ways to accomplish  $\bar{S}$ . There are many  $Z$ 's, each with its own tax vector  $\tau$ , which will produce  $\bar{S}$ .<sup>7</sup> How can we know when we have found an efficient one? At least a partial answer is provided by

**THEOREM III.** *A price equilibrium defined by Definition I, with levels of synthesized pollution  $\bar{S} = \phi(\bar{Z})$ , is efficient, in the sense that there is no Pareto-preferred way to accomplish  $\bar{S}$ , if*

$$\sum_l \bar{\tau}_l \bar{Z}_l \geq \sum_l \bar{\tau}_l Z_l \quad \text{for all } Z \text{ such that} \\ \phi(Z) = \phi(\bar{Z}).$$

<sup>7</sup> As long as  $R < L$ , as assumed. If  $R \geq L$ , then it becomes impossible, in general, to find a set of  $L$  tax-rates which will produce specified levels of  $R$  pollutants; at best, there is no room for choice of  $\tau$ , and hence no question of efficiency. We assume away the problem of more goals than instruments. In any case, if  $R \geq L$ , we may as well stay with the primary pollutants.

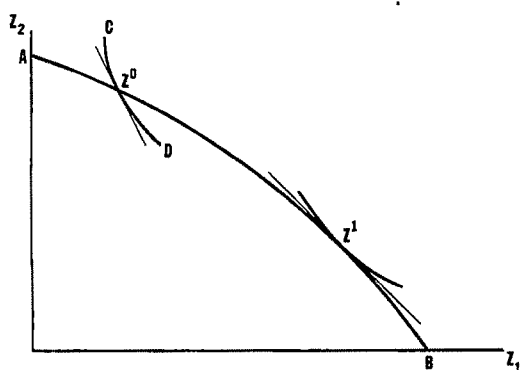


FIGURE 1

The theorem implies that if there is no way to rearrange production so as to increase tax revenue at current rates, without also changing the levels of synthesized pollution, the allocation is efficient.

#### PROOF:

Let  $\bar{q}$ ,  $\bar{x}$ ,  $\bar{z}$  be the allocation and  $\bar{p}$  the price vector for the general competitive equilibrium corresponding to tax vector  $\bar{\tau}$ , and assume the hypothesis of the theorem. Then, if  $\hat{q}$ ,  $\hat{x}$ ,  $\hat{z}$  is any feasible allocation with the same levels of synthesized pollution, i.e.,  $\phi(\sum_k \bar{z}^k) = \phi(\sum_k \hat{z}^k)$ , we know,

$$(a) \quad \sum_l \bar{\tau}_l \bar{z}_l \geq \sum_l \bar{\tau}_l \hat{z}_l$$

Just as in Theorem I, profit maximization and feasibility imply

$$\begin{aligned} \sum_i \bar{p}_i \left[ \sum_k \bar{x}_i^k - \sum_k \hat{x}_i^k \right] \\ \geq \sum_l \bar{\tau}_l \left[ \sum_k \bar{z}_l^k - \sum_k \hat{z}_l^k \right] \end{aligned}$$

Since (a) implies the right-hand side above is nonnegative, we have

$$\sum_i \bar{p}_i \sum_k \bar{x}_i^k \geq \sum_i \bar{p}_i \sum_k \hat{x}_i^k,$$

which is just (a) of Theorem I, from which the rest of the proof follows exactly as in Theorem I.

Theorem III seems to suggest that the pollution control authorities can achieve any level of synthesized pollution efficiently by choosing a set of pollution taxes which maximizes tax revenue, subject to the constraint that smog be at the specified levels. This interpretation is not quite correct, however, primarily because the taxing authority cannot be allowed to exploit its inevitable monopoly position. For example, suppose that in a two-pollutant, single synthesized-product world, the pollution authorities select taxes  $\tau_1^0$  and  $\tau_2^0$ , and in the resulting equilibrium primary pollutant levels are  $Z_1^0, Z_2^0$ , with "smog" level of  $S^0$ , as in Figure 1. The curve  $A-Z^0-Z^1-B$  is an iso-smog curve, and will have the suggested shape if the primary pollutants exhibit "increasing marginal damage," in a generalized sense. The curve  $C-Z^0-D$  is an iso-profit curve for firms; that is, at pollution levels along this curve, firms can produce goods with the same profit for themselves, and can get higher profits only by producing pollution levels above the curve.

Given the tax vector  $\tau^0$ , producers are in equilibrium at  $Z^0$ , which is clearly an inefficient position; points such as  $Z^1$ , for example, have higher profits and hence consumption of greater utility, with the same smog levels. They also have higher tax revenues at the present rates  $\tau^0$ . Therefore, the authorities should encourage movement toward  $Z^1$ , and the obvious way to do this is to lower  $\tau_1$ , relative to  $\tau_2$ , and adjust the absolute levels of both until some point such as  $Z^1$  is reached. At  $Z^1$ , there is no  $Z$  with the same  $\phi(Z)$  and higher tax revenues at the prices  $\tau^1$ . However, whether total tax revenue  $\sum \tau_i^1 Z_i^1$  is greater or less than it was originally,  $\sum \tau_i^0 Z_i^0$ , is impossible to say; this depends on the elasticities of demand for waste disposal services. Hence, Theorem III does *not* imply the control authorities should maximize tax revenues; but it does suggest a tax-adjustment rule they might be able to use to find an efficient solution, and which requires them to know only the  $\phi(Z)$ . And there is still a simple method of estimating the marginal cost of reducing the level of

$S_r$ , given the "marginal product" of primary pollutant  $Z_i$  in the production of  $S_r$ ,  $\partial\phi_r/\partial Z_i$ :

$$\left. \frac{\partial Y}{\partial S_r} \right|_{p=\bar{p}} = \frac{\bar{\tau}_i}{\left( \frac{\partial \phi_r}{\partial Z_i} \right)}$$

All these theorems hypothesize the existence of a competitive price equilibrium; it is possible, of course, that an efficient allocation will not be such an equilibrium, because of nonconvexities in production and/or preferred-to sets. It is also possible, in the case of synthesized "smog," that a price equilibrium will be efficient even though tax revenue (at current rates and subject to the constraint that smog levels be unchanged) is not maximized as hypothesized in Theorem III. Such a situation is illustrated in Figure 2, where smog level  $S^0$  is produced most efficiently, i.e., at highest profits, by pollutant output vector  $Z^0$ , yet  $Z^0$  is on the lowest iso-tax-revenue line which touches the  $S^0$  iso-smog curve; tax revenues at unchanged rates and smog levels are minimized at this efficient point, rather than maximized as hypothesized in Theorem III. The conditions of Theorems I and III are sufficient but not necessary for efficiency.

Another form of difficulty arises if we admit that pollution affects production functions. This external effect is no doubt a diseconomy in the vast majority of cases—smog damages agricultural crops, rots tires, peels paint; but in some instances pollution can be a benefit—oxygen-depleted waters cannot support marine life which attacks piers and boats; some marine life flourishes near sewage or heat outfalls. It is not really known how important such effects are, and it is not unreasonable to assume that they are small relative to the direct effects on consumers; such things as health problems, destroyed aesthetic values, lowered property values, are captured by putting pollution into individual utility functions rather than into production functions, and it is these effects which are the most widely discussed and condemned. But, positive or neg-

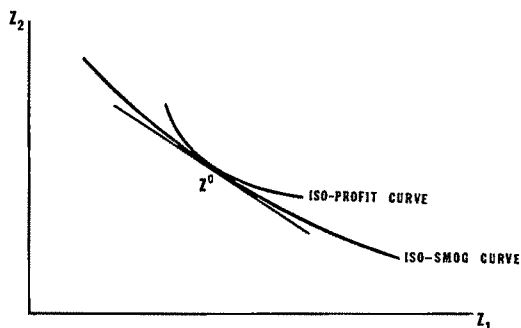


FIGURE 2

ative, significant or negligible, such production-related effects do complicate the analysis.

When we write production constraints in a form which allows these effects, say in the form  $g^k(x^k; z^k; Z) \geq 0$ , or  $g^k(x^k; z^k; S) \geq 0$  for the synthesis case, Theorems I and III still hold;<sup>8</sup> there is no Pareto-preferred way to accomplish the same pollution levels. But now there is the possibility that we can decrease pollution levels and produce more goods, since decreases in pollution levels may make all processes more productive. Of course, if pollution were this bad, any political decision would almost certainly demand its reduction. And, long before we got to a reasonable level of pollution, further reduction would cease to be free.

When pollution levels do affect production directly, the cost of pollution control has two components which must be considered. The first is the value of output which would be lost if each process reduced its pollutant emissions, given the aggregate pollution levels; by Theorem II, this component can be estimated by the tax rates. The second component is the value of the additional resources which would have to be used to produce the same net output of products and pollutants when aggregate pollution levels fall; this component, which presumably is negative, can be estimated only by studying each process directly. For marginal reduc-

<sup>8</sup> Assuming that in choosing their profit-maximizing input-output combinations, firms ignore their individual impact on pollution levels.

tions in pollution, the sum of these components is the net economic cost of pollution reductions, and may be positive or negative.

## REFERENCES

- K. J. Arrow, "An Extension of the Basic Theorems of Classical Welfare Economics," in J. Neyman, ed., *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley 1951, pp. 507-32.
- R. H. Coase, "The Problem of Social Cost," *J. Law Econ.*, Oct. 1960, 3, 1-44.
- A. V. Kneese, *The Economics of Regional Water Quality Management*, Baltimore 1964.
- A. V. Kneese and B. T. Bower, *Managing Water Quality*, Baltimore 1968.
- C. R. Plott, "Externalities and Corrective Taxes," *Economica*, Feb. 1966, 33, 84-87.

# "Fixed Costs" and the Competitive Firm Under Price Uncertainty: Comment

By IRWIN BERNHARDT\*

In a recent article in this *Review*, Agnar Sandmo makes the following statement

One of the basic results in the theory of the firm under certainty is that fixed costs do not matter . . . . This is not so under uncertainty . . . . *Decreasing absolute risk aversion is a necessary and sufficient condition for  $\partial x/\partial B$  [the partial derivative of output with respect to "fixed costs"] to be negative.* [p. 68]

This statement is formally correct in terms of his model as Sandmo neatly proves. But the statement may be misleading.

Sandmo states a "short-run" static model of a price taker. Price,  $p$ , is not known; but there is a subjectively determined probability density function  $f(p)$ . The firm seeks to choose the output per period,  $x$ , that will maximize expected von Neumann-Morgenstern utility. Sandmo writes the utility function with profit per period,  $\pi$ , as its argument. Profit is defined as

$$(1) \quad \pi(x) = px - C(x) - B,$$

where  $C(x)$  "is the variable cost function, and  $B$  is 'fixed cost'" (Sandmo, p. 66). The term  $B$  is a current manifestation of sunk costs,<sup>1</sup> fixed and completely independent of output. Thus  $B$  is the value of an annuity over the relevant number of periods that could have been purchased if the sunk costs had been avoided. Sandmo is perfectly correct to have defined utility,  $U$ , as

$$(2) \quad U = U(px - C(x) - B)$$

Utility might also have been written correctly as

\* Associate professor of management sciences, University of Waterloo.

<sup>1</sup> See Louis De Alessi.

$$(3) \quad U^+ = U^+(px - C(x) + rW - B),$$

or

$$(4) \quad U^- = U^-(px - C(x)),$$

where  $rW$  is the income per period the firm earns from wealth invested elsewhere, and where, for the same values of  $px - C(x)$ ,  $U = U^+ = U^-$ . (Linear transformation of these would also be correct, of course.) The difference among the three representations of utility is that if we hold prices of other goods fixed, equation (2) and equation (4) are *ceteris paribus* relations whereas equation (3) is not. In equation (2),  $U$  is correct as long as  $rW$  is held fixed. In equation (4),  $U^-$  is correct as long as  $rW - B$  is held fixed. The utility a firm has for an increment to income from a given venture depends on its income from other ventures. This is the reason that, as Sandmo shows, if  $B$  were different the choice of  $x$  would be different. It is not  $B$  as fixed costs, qua fixed costs, but  $B$  as a current manifestation of a reduction in wealth that affects output choice. If the firm's managers had, on the morning of the day they chose  $x$ , received word that the firm had unexpectedly won (lost) a tax suit unrelated to this venture, the effect of this on the choice of  $x$  would have been the same as receiving word that sunk costs associated with this venture had been overstated (understated) and a new lower (higher) value was correct.

## REFERENCES

- L. De Alessi, "The Short Run Revisited," *Amer. Econ. Rev.*, June 1967, 57, 450-62.
- A. Sandmo, "On the Theory of the Competitive Firm Under Price Uncertainty," *Amer. Econ. Rev.*, Mar. 1971, 61, 65-73.

# “Fixed Costs” and the Competitive Firm Under Price Uncertainty: Reply

By AGNAR SANDMO\*

Irwin Bernhardt's point concerning the treatment of fixed costs in my article is quite correct. The effect of a change in fixed costs is a pure wealth effect similar to the one analyzed in portfolio theory by Kenneth Arrow (1965, lecture 2; 1970, ch. 3). My analysis may be misleading if one thinks of fixed costs as directly related to the firm's stock of real capital, which is often assumed to be given in short-run analysis. It may therefore be useful to examine the relationship of my analysis to portfolio theory more closely; this is the subject of Section I. Section II presents a reformulation of the model which makes it possible to study the effect of a change in the real capital stock.

## I

To see the relationship to portfolio theory, assume that the argument in the firm's utility function is not profit but final wealth, i.e., its wealth when the price becomes known. If final wealth is  $Y$ , the utility function is accordingly  $U(Y)$  with the properties referred to in my original article.<sup>1</sup>

Let  $A$  be the initial wealth of the firm. This can either be invested in a risk-free asset<sup>2</sup> with a rate of return  $r$  or it can be used to cover the costs of production in the firm, to be denoted now as  $\phi(x)$ . The initial budget constraint is then

$$(1) \quad A = m + \phi(x),$$

where  $m$  is the amount invested in the safe asset. The final wealth is

$$(2) \quad Y = m(1 + r) + px,$$

i.e., the gross return on the safe investment

\* The Norwegian School of Economics and Business Administration.

<sup>1</sup> As in my original article, the underlying preference ordering may be that of a single owner or that of a group of owners whose individual orderings can be aggregated.

<sup>2</sup> This investment might be negative, in which case the firm would borrow to finance its costs.

plus sales revenue. Eliminating  $m$  from (1) and substituting in (2) we then have

$$(3) \quad Y = px - (1 + r)\phi(x) + (1 + r)A$$

Expected utility now becomes

$$(4) \quad E[U(px - (1 + r)\phi(x) + (1 + r)A)]$$

If we now make the purely formal substitutions

$$(5a) \quad (1 + r)\phi(x) = C(x),$$

$$(5b) \quad (1 + r)A = -B,$$

we are back at the model used in my article with final wealth being identified as profit. It is then clear that an increase in  $B$  can be interpreted as a decrease in  $A$ , as noted by Bernhardt. This formulation makes it also quite clear that the rate of return which could be earned on the resources tied up in the firm by investing in the sure asset, should be counted as an element of total costs. This is in fact the point alluded to in footnote 10 of my original article.

## II

We return now to the original formulation of utility as a function of profit. We wish to analyze another type of change in fixed costs in which the basic change is in the volume of fixed equipment owned by the firm. It is perhaps natural to associate this equipment with the firm's stock of real capital. In the short run this stock must be taken as given. It must be assumed, however, that total costs increase with the level of the capital stock and that short-run marginal cost is not invariant to this level.

The cost function can now be written as

$$(6) \quad C = C(x, K),$$

where  $K$  is the capital stock. It is assumed that both partial derivatives  $C_x$  and  $C_K$  are positive. We shall also assume that an in-

crease in capital stock always reduces short-run marginal cost, i.e.,  $C_{xK} < 0$ . The firm now maximizes

$$(7) \quad E[U(px - C(x, K))]$$

with respect to  $x$ . This gives the first-order condition

$$(8) \quad E[U'(\pi)(p - C_x)] = 0,$$

and the second-order condition

$$(9) \quad D = E[U''(\pi)(p - C_x)^2 - U'(\pi)C_{xx}] < 0$$

Differentiating in (8) with respect to  $K$  and solving for  $\partial x / \partial K$  we obtain

$$(10) \quad \frac{\partial x}{\partial K} = \frac{1}{D} C_{xK} E[U''(\pi)(p - C_x)] + \frac{1}{D} C_{xK} E[U'(\pi)]$$

Of the two terms on the right-hand side, the first is analogous to the expression in equation (14) in my original article; it is the

wealth effect<sup>3</sup> of an increase in the capital stock. On the assumption of decreasing absolute risk aversion this effect is negative. The second term is positive, representing the expansionary effect on output of a decrease in marginal cost. If  $C_{xK}$  is very close to zero, the second term can be ignored and we are back at the original analysis. However, it should be stressed that we have here two different problems, and that it is perfectly possible for the wealth effect on output to be negative while the full effect of an increase in the real capital stock is positive.

# REFERENCES

- K. J. Arrow, *Aspects of the Theory of Risk-Bearing*, Helsinki 1965.  
 ———, *Essays in the Theory of Risk-Bearing*, Amsterdam 1970.  
 A. Sandmo, "On the Theory of the Competitive Firm under Price Uncertainty," *Amer. Econ. Rev.*, Mar. 1971, 61, 65-73.

<sup>3</sup> The term "wealth effect" is not entirely appropriate since the reference is actually to a *decrease* in wealth as the consequence of an increase in fixed costs.

# Separability and Complementarity

By EUGENE SILBERBERG\*

In a recent note in this *Review*, H. H. Liebhafsky showed that: (i) The combined assumptions of separability of the utility function and diminishing marginal utility of each good leads to the condition that all goods are "normal" (superior); (ii) If one good, say good 1, exhibits increasing marginal utility (the second-order conditions allow at most one good to have this property), then the remaining  $n-1$  goods must be inferior, while good 1 is normal. The purpose of this note is to provide an extension of that analysis, by showing: (iii) Under the assumptions in (i) above, all goods must be net substitutes; (iv) Under the assumptions in (ii) above, good 1 is a net substitute for goods 2 thru  $n$ , whereas these latter  $n-1$  goods are complementary to each other.

Since the utility function is separable, we shall write it as  $U(x_1, \dots, x_n) = U_1(x_1) + U_2(x_2) + \dots + U_n(x_n)$ . Since only the pure substitution effects are to be analyzed, the problem will be stated as

$$\begin{aligned} \text{Minimize } M &= \sum_{i=1}^n p_i x_i \text{ subject to } U_1(x_1) \\ &+ U_2(x_2) + \dots + U_n(x_n) \\ &= U_0, \end{aligned}$$

where  $U_0$  is a fixed level of utility,  $p_i$  is the unit price of commodity  $x_i$  and  $M$  is the total budget outlay. The cost-minimization problem is usually encountered in the theory of production (see, e.g., Paul Samuelson, ch. 4). It can also be used in consumer theory, since it produces, as first-order conditions, the same tangency condition as the utility maximization problem (ratio of marginal utilities equals ratio of prices), but explicitly keeps the consumer on the same indifference level when prices change (see equations (1) and (2) below.) Hence the comparative statics of this problem must yield the pure substitution terms of the

Slutsky equation (Hicks formulation).

The appropriate lagrangean is

$$L = \sum p_i x_i + \mu(U_0 - \sum U_i(x_i))$$

producing the first-order relations

$$\begin{aligned} (1) \quad \mu U'_i - p_i &= 0, \quad i = 1, \dots, n, \\ (2) \quad U_0 - \sum U_i(x_i) &= 0, \end{aligned}$$

where  $U'_i = \partial U_i / \partial x_i$ . The lagrange multiplier  $\mu$  is the marginal cost of utility. It is the reciprocal of the lagrange multiplier encountered in the utility maximization problem, namely, the marginal utility of money income (see James Henderson and Richard Quandt, p. 52). Assume the second-order conditions hold. (They are, of course, the same as in the utility maximization problem used by Liebhafsky (see Henderson and Quandt, p. 52)). The  $n+1$  equations (1) and (2) can then be solved (in principle) for the "compensated," i.e., real-income-held-constant demand functions

$$(3) \quad x_i = x_i(p_1, \dots, p_n, U_0), \quad i = 1, \dots, n$$

Also produced is an equation for  $\mu$ ,

$$(4) \quad \mu = \mu(p_1, \dots, p_n, U_0)$$

The partial derivatives of (3) with respect to prices thus represent terms of the Hicksian pure substitution matrix.

Differentiating the  $i$ th equation in (1) with respect to  $p_j, j \neq i$ ,

$$(5) \quad \mu U''_{ij} \left( \frac{\partial x_i}{\partial p_j} \right)_{U_0} + U'_i \frac{\partial \mu}{\partial p_j} = 0$$

Using the well-known "reciprocity relation"  $\partial \mu / \partial p_j = \partial x_j / \partial U_0$  (see Samuelson, equation (46), p. 66),<sup>1</sup> equation (5) yields

<sup>1</sup> Samuelson derives this relation in the context of production theory. However, the mathematical structures of his and the present cost-minimization problem are identical. Samuelson's  $v, x, \lambda$ , and  $w$  are my  $x, U_0, \mu$ , and  $p$ , respectively.

\* University of Washington.

$$(6) \quad \left( \frac{\partial x_i}{\partial p_j} \right)_{U_0} = - \frac{U_i'}{\mu U_i''} \left( \frac{\partial x_j}{\partial U_0} \right)$$

Using Liebhafsky's results ((i) and (ii) above), (iii) and (iv) follow immediately:

(iii) If  $U_i'' < 0$  for all  $i = 1, \dots, n$ , then all goods are superior<sup>2</sup> ( $\partial x_j / \partial U_0 > 0$ ),  $j = 1, \dots, n$ , and hence  $(\partial x_i / \partial p_j)_{U_0} > 0$ ,  $i, j = 1, \dots, n$ .

(iv) If  $U_i'' > 0$ , then  $\partial x_j / \partial U_0 < 0$ , hence  $(\partial x_i / \partial p_j)_{U_0} = (\partial x_j / \partial p_i)_{U_0} > 0$ . If  $U_k'' > 0$ ,  $k \neq i, j$ ,  $\partial x_j / \partial U_0 < 0$  (the remaining goods are inferior) and hence  $(\partial x_i / \partial p_j)_{U_0} < 0$ .

<sup>2</sup>  $\partial x_i / \partial U_0$  and  $\partial x_j / \partial M$  (from the utility maximization problem) can differ only by units, not by sign, as long as increases in money income yield increases in utility, prices held constant. With constant prices, changes in  $U_0$  in the cost-minimization problem must trace out precisely the income-consumption paths generated by changes in  $M$  in the utility maximization scheme. Hence, superiority or inferiority of the goods can be defined equivalently either in terms of the sign of  $\partial x_j / \partial U_0$  or the sign of  $\partial x_j / \partial M$ .

There do not appear to be similar theorems concerning the "gross substitution" effects, i.e., those containing income effects. We see, however, that separability of the utility function places very severe restrictions on the properties of the resulting demand functions: Either all commodities are superior and net substitutes, or, only one commodity is superior and the rest inferior, with the superior commodity a substitute for the inferior ones, and the inferior commodities all net complements to each other.

#### REFERENCES

- J. M. Henderson and R. E. Quandt, *Microeconomic Theory*, New York 1958.  
 H. H. Liebhafsky, "New Thoughts About Inferior Goods," *Amer. Econ. Rev.*, Dec. 1969, 59, 931-34.  
 P. A. Samuelson, *Foundations of Economic Analysis*, Cambridge, Mass. 1947.

# Money Illusion and the Aggregate Consumption Function: Comment

By ALEX CUKIERMAN\*

In their article in this *Review*, William Branson and Alvin Klevorick test for the existence of money illusion in the U.S. economy by using a regression of real consumption on real income, wealth, and an aggregative index of the consumer price index where all the regressors are lagged several periods.

Without introducing the lags, the basic structure of the model is in their notation (equation (4), p. 834):

$$\log c_t = b_0 + b_1 \log y_t + b_2 \log w_t + b_3 \log P_t$$

where  $c_t$  is real consumption,  $y_t$  real net labor income,  $w_t$  real consumer net worth, and  $P_t$  is the price level of consumer goods.

The test for the existence of money illusion is formulated on the coefficient  $b_3$ . It is claimed that if there is no money illusion, real consumption  $c_t$  will be pushed only by  $y_t$  and  $w_t$ . If there is money illusion in the Patinkin sense,  $b_3$  will be significantly different from zero and positive because when prices and consequently nominal incomes rise, people feel themselves richer and are motivated to spend on consumption more than they would have, had the price level remained constant. Hence they claim if there is money illusion,  $b_3 > 0$ .

This formulation of the money illusion test depends implicitly on the theoretical assumption that when  $P$  moves all its components (namely the price  $P_i$  of the particular goods which enter into the index  $P$ ) move equiproportionally. No money illusion will show up as no change in real consumption when  $P$  moves up *only* when this equiproportional movement in all prices is there. But as soon as one recognizes that the real world does not provide us with such laboratory controlled experiments, one will have

to take into account the various substitution effects as a result of changes in relative prices which accompany any movement in  $P$ . In what follows this will be done and the relationship between  $P$  and  $P_i$  as it bears on the test of money illusion will be investigated.

It will be shown in particular that if one admits that the "true" formulation of the consumption function is in terms of particular prices rather than in terms of one aggregate index (thus allowing substitution effects to show in consumption), then:

1) the test of existence of money illusion cannot in general be formulated on  $b_3$ . Only under very special circumstances will the nonexistence of money illusion imply  $b_3 = 0$ ;

2) the inconsistencies in the estimators of  $b_1$ ,  $b_2$ , and  $b_3$  will be derived. It will be shown in particular that the *Plim* of the estimator of  $b_3$  is not independent of the individual prices ( $P_i$ );

3) the test of money illusion is redone in Section III with five individual price indices instead of the single *CPI*. The original conclusion of the authors concerning the existence of money illusion is weakened but not altered.

## I. When Will $b_3 = 0$ Be Equivalent to No Money Illusion?

Let real consumption  $c$  be

$$(1) \quad c = \frac{C}{P} = \frac{\sum P_i C_i}{P}$$

where  $C_i$  is the quantity of the  $i$ th good consumed in the economy,  $C$  is consumption in current prices, and  $P$  is a price index which converts current consumption to fixed base prices and is defined by the relation

$$(2) \quad P = \sum W_j P_j$$

\* Ph.D. candidate in economics at M.I.T. I would like to thank Franklin Fisher and Edwin Kuh for helpful comments.

where  $W_j$  are some weights such that<sup>1</sup>  $\sum W_j = 1$ .

Let  $\beta_1$ ,  $\beta_2$ , and  $\gamma$  be, respectively,<sup>2</sup> the coefficients of real income, wealth, and the price level index in the consumption equation. Then the Branson and Klevorick consumption equation is

$$(3) \quad c_t = e^{\beta_0} y_t^{\beta_1} w_t^{\beta_2} P_t^\gamma e^{\epsilon_t}$$

where  $\epsilon_t$  is an error term. To allow for change in relative prices, one will have to consider the true model

$$(4) \quad c_t = e^{\beta_0} y_t^{\beta_1} w_t^{\beta_2} P_1^{\gamma_1} \dots P_n^{\gamma_n} e^{\epsilon_t}$$

In this alternative formulation price relatives are free to change. But after having estimated the coefficients one can *impose* an equiproportionate change in prices and test whether consumption is thereby affected or not. Namely

$$H_0; \text{ no money illusion} \Leftrightarrow \sum_{i=1}^n \gamma_i = 0$$

A

$$H_1; \text{ money illusion} \Leftrightarrow \sum_{i=1}^n \gamma_i > 0$$

The corresponding hypotheses as formulated by Branson and Klevorick are

$$H_0; \text{ no money illusion} \Leftrightarrow \gamma = 0$$

B

$$H_1; \text{ money illusion} \Leftrightarrow \gamma > 0$$

In what follows the relationship between  $\gamma$  and  $\sum \gamma_i$  will be worked out and used in order to show that  $\sum \gamma_i$  does not necessarily imply  $\gamma = 0$ . Note that by definition of elasticity

$$(5) \quad \gamma_j = \frac{P_j}{c} \frac{\partial c}{\partial P_j} \quad j = \dots n$$

Define

$$(6) \quad \eta_j = \frac{P}{P_j} \frac{\partial P_j}{\partial P} \quad j = 1 \dots n$$

<sup>1</sup> For the Laspeyre price index which is used by Branson and Klevorick they are the base period weights. But the analysis goes through for shifting weight base indices. See footnote 6.

<sup>2</sup>  $\beta_1$ ,  $\beta_2$ , and  $\gamma$  are a new notation for Branson and Klevorick's  $b_1$ ,  $b_2$ , and  $b_3$ , respectively.

Again by the definition of elasticity

$$(7) \quad \gamma = \frac{P}{c} \frac{\partial c}{\partial P} = \frac{P}{c} \sum_{j=1}^n \frac{\partial c}{\partial P_j} \frac{\partial P_j}{\partial P} \\ = \sum_{j=1}^n \frac{P_j}{c} \frac{\partial c}{\partial P_j} \frac{P}{P_j} \frac{\partial P_j}{\partial P} = \sum_{j=1}^n \gamma_j \eta_j$$

from which it follows<sup>3</sup>

$$(8) \quad \gamma = \sum_{j=1}^n \gamma_j \eta_j$$

Under the null hypothesis of the true model  $\sum \gamma_j = 0$ . It can be seen from (8) that this does not in general imply  $\gamma = 0$  which is the Branson and Klevorick null hypothesis. Whether it does or does not imply  $\gamma = 0$  will depend on the distribution of  $\eta_j$  relative to the distribution of  $\gamma_j$ .

To make this point more apparent it will be worthwhile to ponder about the possible values of  $\sum \gamma_j \eta_j$  under  $H_0$ .

Since  $\sum \gamma_j = 0$  some of the  $\gamma_j$  will be positive and others will be negative. The coefficients  $\eta_j$  are the price movements in each of the prices  $P_j$  relative to the movement of the general price level. Hence:

- $\eta_j > 1$  for prices which increase more than the average level of prices,
- $\eta_j = 1$  for prices which increase like the average level of prices,
- $\eta_j < 1$  for prices which increase less than the average level of prices.<sup>4</sup>

As long as most prices tend to increase when  $P$  increases  $\eta_j$  will be positive.<sup>5</sup> Hence the term  $\sum \gamma_j \eta_j$  will be composed of positive and negative terms. When will these terms cancel each other to 0 if  $A:H_0$  holds? To see when, note

$$\sum_j \gamma_j \eta_j = \eta' \gamma$$

<sup>3</sup>  $\eta_j$  is an *ex post* measured quantity and does not reflect a functional relationship between  $P_j$  and  $P$  but it may nevertheless display a certain pattern.

<sup>4</sup>  $\eta_j < 0$  for prices which decrease when the average level increases.

<sup>5</sup> It does seem reasonable to assume that in inflationary periods  $\eta_j > 0$ . In any case it can be tested empirically.

where  $\eta' = (\eta_1 \dots \eta_n)$   
by

$$A:H_0 \quad (1 \dots 1)\underline{\gamma} = 0 \Rightarrow \lambda(1 \dots 1)\underline{\gamma} = 0$$

( $\lambda$  any scalar)

Hence if  $\eta_j = \text{const.}$   $j=1 \dots n \Rightarrow \eta'\gamma = 0$ . But  $\eta_j = \text{const.}$   $j=1 \dots n$  means that all prices have increased equiproportionally. Hence if such a happy case happens, "No money illusion" implies  $\gamma = 0$  and the existence of money illusion implies  $\gamma > 0$  (for  $\text{const.} > 0$ ). As a matter of fact in this case  $\gamma = \Sigma \gamma_j$  since if all individual prices change relative to the average price level in the same proportion the average price level which is a weighted average of those prices will also increase in this same proportion. It follows  $\eta_j = 1$ ,  $j=1 \dots n$ .

But as can be seen this is a very particular case. The coefficients  $\eta_j$  and  $\gamma_j$  may be of equal signs. Then  $\Sigma \eta_j \gamma_j$  will be positive even if  $\Sigma \gamma_j = 0$ , namely even if there is no money illusion. Conversely if  $\eta_j$  and  $\gamma_j$  have opposite signs for most  $j$ ,  $\Sigma \eta_j \gamma_j$  will be negative even if  $\Sigma \gamma_j = 0$ .

Branson and Klevorick find that  $\gamma > 0$  and conclude from it that *there is* money illusion. But as the above analysis shows this conclusion is unwarranted since no money illusion is consistent with  $\gamma > 0$ . If for instance there is a positive sign correlation between  $\eta_j$  and  $\gamma_j$ ,  $\gamma$  will be positive even if there is no money illusion. Whether this is so or not is an empirical question to be investigated.<sup>6</sup>

## II. The Estimation Bias as a Result of Misspecification

We saw in the former section that the formulation of  $H_0$  in the form  $\gamma = 0$  will usually cause the test of money illusion to be either wrong or inconclusive.

But in addition there will be in general a bias in the estimators of the coefficients of equation (3) since it is a misspecified version of equation (4).

In what follows the bias expression will be

<sup>6</sup> Branson and Klevorick also use a shifting weight base price index. This does not change the relationship between  $\gamma$  and  $\Sigma \gamma_j$  in equation (8) and hence does not change the basic conclusion. Namely  $\Sigma \gamma_j = 0$  does not necessarily imply  $\gamma = 0$ .

derived. It will be shown that there will in general be a bias in the estimators of all the coefficients of equation (3) and in particular in the estimator of  $\gamma$ . This latter bias will vanish for large samples only in the case of equiproportional price increase.

Let  $C_{iy}$ ,  $C_{iw}$ ,  $C_{ip}$ , respectively, be the *Plim* of the regression coefficients of the *logs* of real income, real wealth, and the average price level in the regression of  $P_i$  on these variables. Let

$$B = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\gamma} \end{bmatrix}$$

be the vector of estimators which is obtained in the Branson and Klevorick version of the model. Using the Theil specification error theorem<sup>7</sup>

$$(9) \quad \text{Plim } B = \begin{bmatrix} 1 & 0 & C_{1y} & \dots & C_{ny} \\ 0 & 1 & C_{1w} & \dots & C_{nw} \\ 0 & 0 & C_{1P} & \dots & C_{nP} \\ & & 3 \times (n+2) & & \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \gamma_1 \\ \vdots \\ \gamma_n \end{bmatrix}$$

The regression of  $y$  on itself,  $w$ , and  $P$  results in a coefficient of 1 for  $y$  and 0 for the other regressors.

Similarly the regression of  $w$  on  $y$ ,  $w$ , and  $P$  yields a coefficient of 1 to  $w$  and 0 to the other regressors. This is the reason for the appearance of the 0 and 1 in the first two columns of the matrix on the right-hand side of (9).

*What is  $\hat{\gamma}$  a Consistent Estimator of?*

From (9) it follows<sup>8</sup>

$$(10) \quad \text{Plim } \hat{\gamma} = \sum_{j=1}^n \gamma_j C_{jP}$$

<sup>7</sup> See Henri Theil, ch. 6.2.4, pp. 211-15, and Appendix 6B, p. 327.

<sup>8</sup> It also follows from equation (9) that

$$\text{Plim } \hat{\beta}_1 = \beta_1 + \sum_{j=1}^n C_{jy} \gamma_j \quad \text{and} \quad \text{Plim } \hat{\beta}_2 = \beta_2 + \sum_{j=1}^n C_{jw} \gamma_j$$

Hence  $\hat{\beta}_1$  and  $\hat{\beta}_2$  will in general be inconsistent estimators of  $\beta_1$  and  $\beta_2$ . The direction of the inconsistencies will again depend on whether there is or there is not some systematic relationship between  $C_{jy}$ ,  $C_{jw}$ , and  $\gamma_j$ .

But since  $C_{jP}$  is the coefficient of the logarithm of the average price level in the regression of the logarithms of the  $j$ th price on the logarithms of real income, real wealth, and the average price level, it follows that in the notation of Section I

$$(11) \quad C_{jP} = \frac{\partial \log P_j}{\partial \log P} = \eta_j$$

Hence,

$$(12) \quad \text{Plim } \hat{\gamma} = \Sigma \gamma_j \eta_j = \gamma$$

Namely  $\text{Plim } \hat{\gamma}$  is a consistent estimator not of the "true" coefficient of  $P$  in equation (4) (which is 0) but of  $\gamma$  which is a weighted average of the  $\gamma_j$ . In a sense it is comforting for the Branson and Klevorick way of testing for money illusion since it shows that the  $\gamma$  they intended to estimate is estimated consistently by  $\hat{\gamma}$ . But on the other hand it wipes out the possibility that the error in specification might by a happy coincidence outbalance the error committed in testing money illusion on  $\gamma$  rather than on  $\Sigma \gamma_j$ . Considering that the criticism of Section I has even more weight since the validity of the Branson and Klevorick test depends again on whether

$$\Sigma \gamma_i = 0 \Rightarrow \Sigma \gamma_i \eta_i = 0$$

and

$$\Sigma \gamma_i > 0 \Rightarrow \Sigma \gamma_i \eta_i > 0$$

The answer to that problem will again depend on the kind of relationship that might exist between  $\gamma_j$  and  $\eta_j$ . It certainly can't be answered a priori and would require an empirical investigation. But it seems that rather than investigate empirically whether the conditions needed for the validity of the Branson and Klevorick test are fulfilled or not it might be more fruitful to test money illusion on the "true" model, to start with. Such a test is described in the next section.

### III. Reestimation of the Money-Illusion Consumption Function

The main target of the following analysis is to introduce individual prices into the money-illusion consumption function and

test whether the result obtained by Branson and Klevorick is thereby substantially changed. As a by-product it will be possible to get some notion of the influence of particular prices on aggregate real consumption and to test the results against the predictions provided by consumer theory.

The general lag specification of the equation to be estimated is given in Branson and Klevorick's notation by<sup>9</sup>

$$(13) \quad \log c_t = b_0 + \sum_{i=0}^I \log y_{t-i}^{\gamma_i} + \sum_{j=0}^J \log w_{t-j}^{\delta_j} + \sum_{s=0}^{T_1} \log P_{1,t-s}^{\eta_s^1} + \dots + \sum_{s=0}^{T_n} \log P_{n,t-s}^{\eta_s^n}$$

The first rather arbitrary decision that one must make relates to the degree of disaggregation of the price variable. Ideally, the more disaggregation the better will be the model specification, since all possible substitution effects are going to be taken into account. But there are several difficulties mainly of statistical nature with such a corner solution. First, the more disaggregated the price variable the higher will be the likelihood of multicollinearity because of the tendency of prices within certain groups of consumption goods to move fairly uniformly over time. This would tend to remove any reliability from the coefficients of individual prices.

Second, data at a thoroughly disaggregated level are usually less reliable since the proportion of random noise in it is higher. Hence, as in many economic problems, the optimum (or compromise) solution must be somewhat in the middle. I have chosen to disaggregate Branson's and Klevorick's  $P$  into the five broad categories of the *CPI* reported in the *Survey of Current Business*.<sup>10</sup> These are:

- $P_A$  Price index of Apparel
- $P_F$  Price index of Food
- $P_H$  Price index of Housing

<sup>9</sup> The superscript on  $\eta$  indicates the price category.

<sup>10</sup> See the Appendix for further details.

- $P_R$  Price index of Recreation  
 $P_T$  Price index of Transportation

The rest of the data was kindly provided by Branson and Klevorick, including:

- $c_t$  real consumption per capita  
 $y_t$  real net labor income per capita  
 $w_t$  real consumer net worth

As a result the Branson and Klevorick results and those presented here are directly comparable since they refer to the same quarterly data and span the same period—from the first quarter of 1953 to the last quarter of 1965.

The estimation to be presented here can be viewed as an exercise in appraising the effects of disaggregation of the price variable on the test of money illusion and on the values of the coefficients of the real wage and net real wealth variables.

If the results change a lot, the inescapable conclusion would be that one has to work with disaggregated data to get meaningful results. If they do not change significantly more weight could be assigned to aggregated studies on the consumption function, since they capture all the essential elements that can be obtained in a disaggregated model.

As in all single equation models the equation estimated is in reality part of a simultaneous system. As a result the single equation estimators of the coefficients in equation (13) might be biased. I have not tried to account for that but if one can judge from the attempt done by Branson and Klevorick on their aggregated equation to account for simultaneity this doesn't affect the coefficients significantly (see p. 845–46).

A second arbitrary decision refers to the length of the lags to be used for each of the independent variables in equation (13). This required some experimentation with alternative length of lag specifications. The criterion for determining the final equation to be picked as representative of the truth was the significance of the  $t$ -test pertaining to the lagged variables' coefficients. Sometimes as in the case of the real wage rate the cut off was provided by the lag  $i$  for which  $\gamma_i$  became the first time insignificantly different

from 0 irrespective of whether higher lags had or had not significantly negative coefficients.<sup>11</sup>

The final lag specification provided by Branson and Klevorick (see p. 839, equation 1–4) seems a good place to start but because of multicollinearity it is practically impossible to have a very long lag on *each one* of the five price variables.

There is also ground to believe a priori that the effect of lagged individual prices on consumption will die out quicker than that of the lagged general *CPI*. Since the individual prices are the components of the *CPI* the influence of the last will at least stretch as far back as the influence of each of the individual components. Moreover it might stretch even more into the past if beyond a certain point each lagged price component alone doesn't push consumption significantly but their sum total does.

Consequently the lag specification that was chosen was: A 7-quarter lag on real wages ( $I=7$ ). A 1-quarter lag on real net worth ( $J=1$ ) and a uniform 2-period lag on all the price variables ( $T_A=T_F=T_H=T_R=T_T=2$ ). These preliminary results are reproduced with Branson and Klevorick's main result in Table 1. Even though equation 1–2 is not the final one that will be picked to represent the results it is still interesting to make a first comparison between the Branson and Klevorick results and the disaggregated version 1–2. The overall fit is roughly the same or maybe slightly better as measured by the standard error of the regression.

The main result is that even though the total sum of the price coefficients is amazingly near to the Branson and Klevorick result (0.414 versus 0.418), it is not significantly different from 0, the standard error of the sum being 0.406. Except for the sum of food price coefficients none of the individual sums of price coefficients are significantly different from zero. The net wealth coefficients sum is

<sup>11</sup> Some justification for this procedure can be found in the permanent income hypothesis. Any increase in real wage in the past either increases or doesn't change permanent income. As a result any of the coefficients  $\gamma_i$  will be either positive or 0 but not negative.

TABLE 1—ESTIMATION OF AGGREGATED VERSUS  
DISAGGREGATED MONEY-ILLUSION  
CONSUMPTION FUNCTION<sup>c</sup>

	Equations		
	1-1 <sup>a</sup>	1-2	1-3 <sup>b</sup>
Constant	-1.953	0.699	0.042
Independent Variables	(0.114)	(0.826)	(0.538)
$\ln y$			
$I$	6	7	3
$\Sigma \gamma_i$	0.661	0.550	0.599
	(0.043)	(0.144)	(0.050)
$\ln w$			
$J$	0	1	0
$\Sigma \delta_j$	0.127	0.088	0.136
	(0.036)	(0.060)	(0.039)
$\ln P_A$			
$T_A$		2	1
$\Sigma \eta_s^A$		-0.130	-0.482
		(0.422)	(0.180)
$\ln P_F$			
$T_F$		2	2
$\Sigma \eta_s^F$		0.431	0.202
		(0.203)	(0.066)
$\ln P_H$			
$T_H$		2	2
$\Sigma \eta_s^H$		-0.160	0.700
		(0.395)	(0.166)
$\ln P_R$			
$T_R$		2	2
$\Sigma \eta_s^R$		0.360	0.146
		(0.226)	(0.141)
$\ln P_T$			
$T_T$		2	0
$\Sigma \eta_s^T$		-0.086	-0.184
		(0.096)	(0.060)
$T$	6		
$\Sigma \Sigma \eta_i$	0.418	0.414	0.382
	(0.036)	(0.406)	(0.1342)
$R^2$	0.998	0.999	0.998
$S.E. 10^2$	.2964	0.2810	.3190
$D.W.$	1.757	2.76	1.88

Source: Branson and Klevorick, p. 839, Table 1, equation 1-4. Their equation has all lengths of lags longer by 1 because they include the current period in the length of lag specification while I exclude it.

<sup>a</sup> Branson and Klevorick aggregated equation.

<sup>b</sup> Final price disaggregated equation.

<sup>c</sup> Standard errors in parentheses.

smaller and insignificant now. The real wage rate coefficients sum is slightly smaller but has a much higher variance than in the aggregated equation.

In general one notes that the coefficients of equation 1-2 are of the same order of magnitude as those of equation 1-1 but they have uniformly higher standard deviations. One obvious reason is the higher degree of multicollinearity that is introduced by disaggregation of the price variable. Another reason could be the misspecification of the length of the lags of the different variables.

Since changes in the length of the lags could affect both the values and the reliability of the coefficients it was deemed necessary to experiment with different length of lags for the different variables. Several runs of regressions not reproduced here as well as the distribution of the individual coefficients of  $y$  suggested that the lag on real wages is too long.<sup>12</sup>

Inspection of the other lags suggest that one could do away with the lag on net real wealth ( $w$ ) altogether since it was very small and insignificant compared to the current coefficient of  $w$ .

Using the  $t$ -statistic as a criterion some of the individual price lags seemed too long and others too short. When correction for that was made the same criterion suggested changing the lags again. After experimenting with a 3-period lag on  $y$ , no lag on  $w$ , and varying lags on the different prices, the equa-

<sup>12</sup> The lag structure of the  $\gamma_i$  for equation 1-2 was as follows:

$i$	$\gamma_i$	standard error	$t$ -statistic
0	0.393	0.135	2.91
1	0.037	0.140	0.27
2	0.225	0.138	1.63
3	-0.110	0.121	-0.92
4	0.069	0.124	0.55
5	0.022	0.117	0.18
6	-0.037	0.123	-0.30
7	-0.048	0.124	-0.39

Beyond a 3-period lag the coefficients are relatively very small and insignificant. Some of them are negative, a fact which seems to contradict economic theory. Hence a 3-period lag on the real wage variable seems a better length of lag. Also a different auxiliary regression where all individual prices appeared with no lags at all, net real wealth with 1-period lag and the real wage in a polynomial distributed lag on 7 periods suggested the same conclusion. The distributed lag on wages displayed a classic Koyck distribution of coefficients which converged to 0 beyond the third-quarter lag.

tion labelled 1-3 in Table 1 was picked as representative of the results.

The other equations as well as this last one are reproduced in Table 2, with equation 1-3 of Table 1 shown as 2-3 in Table 2.

There are several features which are common to all those equations:

1) The sums of all the price coefficients, which measure money illusion, remain roughly in the same range as the original

Branson and Klevorick single money-illusion coefficient. As a matter of fact, excluding equation 2-1, all the other equations yield smaller values for the sum of the price coefficients. The range varies between 0.322 and 0.414 (excluding equation 2-1) versus a coefficient of 0.418 for the Branson and Klevorick aggregated equation. More significant is the fact that the standard errors of the sum  $\sum \eta_s^i$  are uniformly higher than

TABLE 2—ESTIMATION OF THE DISAGGREGATED MONEY-ILLUSION CONSUMPTION FUNCTION<sup>a</sup>

	2-1	Equations 2-2	2-3	2-4
Constant	-0.530	0.001	0.042	0.303
Independent Variables	(0.550)	(0.564)	(0.538)	(0.525)
$\ln y$				
$I$	3	3	3	3
$\Sigma \gamma_i$	0.529	0.584	0.599	0.606
	(0.050)	(0.053)	(0.050)	(0.056)
$\ln w$				
$J$	0	0	0	0
$\Sigma \delta_j$	0.184	0.162	0.136	0.133
	(0.039)	(0.038)	(0.039)	(0.044)
$\ln P_A$				
$T_A$	0	1	1	2
$\Sigma \eta_s^A$	-0.111	-0.352	-0.482	-0.577
	(0.161)	(0.184)	(0.180)	(0.187)
$\ln P_F$				
$T_F$	2	1	2	3
$\Sigma \eta_s^F$	0.148	0.108	0.202	0.206
	(0.067)	(0.050)	(0.066)	(0.077)
$\ln P_H$				
$T_H$	1	1	2	2
$\Sigma \eta_s^H$	0.606	0.622	0.700	0.688
	(0.176)	(0.170)	(0.166)	(0.212)
$\ln P_R$				
$T_R$	2	1	2	1
$\Sigma \eta_s^R$	0.006	0.090	0.146	0.179
	(0.142)	(0.150)	(0.141)	(0.139)
$\ln P_T$				
$T_T$	0	1	0	0
$\Sigma \eta_s^T$	-0.125	-0.101	-0.184	-0.174
	(0.062)	(0.068)	(0.060)	(0.062)
$\Sigma \Sigma \eta_s^i$	0.523	0.367	0.382	0.322
	(0.349)	(0.181)	(0.134)	(0.203)
$R^2$	0.998	0.998	0.998	0.998
$S.E. 10^2$	.3630	.3400	.3190	.3120
$D.W.$	1.6	1.86	1.88	1.79

<sup>a</sup> Standard errors in parentheses.

in the Branson and Klevorick formulation. In some cases as in equations 1-2, 2-1, 2-4, the sum of the price coefficient is not even twice its standard error which makes it not significantly different than 0 at the 0.05 level of significance. In the other equations which did yield a result significantly different than 0, the  $t$ -statistic barely reaches a value of 3 (equation 2-3) versus a value of more than 10 in the Branson and Klevorick aggregated version (equation 1-1).

It seems therefore that the conclusion that a sizable degree of money illusion had existed in the U.S. consumption function during the years 1953-65 has less statistical foundation than the aggregated experiment would seem to indicate.

2) With regard to the coefficients of wealth and real wages, the disaggregated versions of Table 2 yield slightly higher coefficients for the wealth variable and slightly lower coefficients for the real wage. For both these coefficients the level of significance remains roughly the same as in the aggregated version.

3) With regard to the particular price lagged coefficient sums, both the food and housing price coefficients are positive and significantly different than 0 at the 0.01 level of significance. The sum of the transportation price index is negative and significant in most equations in Table 2 at the 0.05 level.

Finally the apparel and recreation coefficients sums are not significantly different than 0 even though the first is persistently negative while the second is always positive.<sup>13</sup>

As can be seen from Table 2 the major results are going to be the same whichever equation in Table 2 is picked as final. Equation 2-3 in Table 2 was picked as a representative equation, because it has two desirable properties: the  $D.W.$  statistic is closest to 2, and it has almost as good an  $R^2$  and a stan-

dard error of regression as any other equation. It is also the equation most favorable to the Branson and Klevorick result. Even so it has a  $t$ -statistic for the sum of the price coefficients of only 2.85 versus the very highly significant 11.6  $t$ -statistic that was obtained in the Branson and Klevorick aggregated version.

#### IV. Interpretation of the Results and Conclusion

Reestimation of the money-illusion consumption function seems to indicate that the money-illusion coefficient is smaller and less significant when the consumer price index is disaggregated. One might conjecture that with further disaggregation of prices the money-illusion coefficient could disappear altogether. Nevertheless the results still seem to indicate some degree of money illusion, and aggregation of the price index does not lead to an erroneous conclusion.

At a more fundamental level there are two other factors which could generate money illusion at the aggregate level even if it did not exist for each consumer. One is that the same consumption deflator  $P$  is used for all consumption categories, thus creating some biases in measuring the changes in the real quantities of each consumption category. A second one is caused by the distribution effects which are created as a result of a change in any of the  $P_j$ . Consequently even if  $\sum_j \sum_s \eta^j$  turns out to be different than 0, there is no way of knowing whether the reason for that is "genuine money illusion" or other effects as aggregation error in the price deflator or distribution effects generated by a change in  $P_j$ . If one doesn't care for the distinction and would assign the name of money illusion to any observed relationship between prices and consumption whether it arises because of distribution or other effects, the results become much weaker. If one believes that the main reason for money illusion is from distribution effects it would be more interesting to parameterize the coefficients of the distributional process. For example one could divide income recipients by income classes and estimate the effect of a change in prices on their relative

<sup>13</sup> The sum of the lagged coefficients of a particular price index is the steady state influence of an increase in this particular price on the level of aggregate real consumption. For example, if as in equation 2-3  $\sum \eta^H = 0.7$ , it means that if the price of housing goes up by one percentage point in all periods the aggregate real level of consumption in all periods is going to rise by 0.7 of 1 percent.

shares of national income. Money illusion then might become a combination of those terms with the propensities of each income group to spend on the different consumption categories. But this is beyond the scope of this paper.

#### APPENDIX

##### *Listing of the Data*

All the data is quarterly data between the first quarter of 1953 and the last quarter of 1965. Because of the lags in the model, the consumption observations begin 1955 I. This explains why Branson and Klevorick speak of the sample period running from 1955 I to 1965 IV.

The table in following column lists the basic data used in the study.

#### REFERENCES

- W. H. Branson and A. K. Klevorick, "Money Illusion and the Aggregate Consumption Function," *Amer. Econ. Rev.*, Dec. 1969, 59, 832-43.
- H. Theil, *Economic Forecasts and Policy*, 2nd ed., Amsterdam 1965.

Series	Units	Source
Aggregate real consumption	billions of 1958 \$	Harold Shapiro (unpublished)
Real net labor income	Same	Same
Real consumer net worth	Same	FRB-MIT model
Population	Millions	Same
CPI: Price of Housing	1957-59 = 100	1967 Statistical Supplement to the <i>Survey of Current Business</i>
CPI: Price of Apparel and Upkeep	Same	Same
CPI: Price of Food	Same	Same
CPI: Price of Health and Recreation	Same	Same
CPI: Price of Transportation	Same	Same

*Notes:* Items 1-4 have been provided by the courtesy of Branson and Klevorick; Items 5-9 appear in the Supplement to the *Survey of Current Business* as monthly observations. They have been converted into quarterly data by averaging the monthly observations over each quarter.

# Money Illusion and the Aggregate Consumption Function: Reply

By WILLIAM H. BRANSON AND ALVIN K. KLEVORICK\*

In his comment, Alex Cukierman argues that to obtain better estimates of price, or money-illusion, effects in an aggregate consumption function one should disaggregate the consumer price index (*CPI*) into its components and include these separate price components in the equation, rather than just including the *CPI* as we did. He then estimates a consumption function for our sample period, 1955 I–1965 IV, using our data for real per capita consumption, net labor income, and wealth, and disaggregated data on five individual price series—the *CPI* components for food, housing, apparel, transportation, and health and recreation.<sup>1</sup> In his representative equation, the lag on income is shortened from seven quarters to four quarters while the lags on the individual prices vary from one to three quarters, as compared with our original seven-quarter lag on the *CPI*.<sup>2</sup>

As is clear from Cukierman's Table 1, the coefficient sums of Cukierman's best equation (his 1–3) are fairly similar to those of our final equation (his 1–1). The sums of his income coefficients and price coefficients are a bit smaller than ours, and his wealth coefficient is a bit larger. The main difference between equations 1-1 and 1-3 in Cukierman's Table 1 is that the sum of his price coefficients (in 1-3) is only 2.85 times its standard error, while ours (in 1-1) is 11.6 times its standard error. From this result, Cukierman concludes that the money-illusion

coefficient is smaller and less significant when the consumer price index is disaggregated. Nevertheless the results still seem to indicate some degree of money illusion.

The procedure Cukierman uses raises two questions that are best handled sequentially. First, to what extent are his estimates the result of changing the lag lengths in the estimated equation, and to what extent are they due to disaggregation of the price variable? Second, if disaggregation is the important cause of the divergence between his results and ours, what is the best way to interpret his results? The first two sections below consider these two questions, and the third section concludes with some further comments.

## I. The Effects of Disaggregation on the Estimates

In an attempt to separate the effects of changing lag specification—specifically, shortening lag length—from the effects of disaggregation of the price variable, we have reestimated our equation using the aggregate *CPI* but using Cukierman's lag lengths. The results are shown in Table 1. The table shows several estimates of the parameters of the consumption function,

$$\ln c_t = \beta_0 + \sum_{i=0}^I \gamma_i \ln y_{t-i} + \sum_{j=0}^J \delta_j \ln w_{t-j} + \sum_{k=0}^K \eta_k \ln P_{t-k} + \epsilon_t \quad (1)$$

The data are those used by ourselves and Cukierman, described in our data appendix, and the equations are estimated over the period 1955 I–65 IV. The format of Table 1 is the same as that of our original Table 1 and Cukierman's Table 1.

Equation 1-1 simply reproduces the final estimate of the money-illusion consumption

\* Associate professors of economics, Princeton and Yale Universities, respectively.

<sup>1</sup> Cukierman refers to a *CPI* component for "recreation," by which we assume he means "health and recreation." This is the component of the *CPI* that, taken with the other categories he includes, sums to the aggregate *CPI*.

<sup>2</sup> Note that in our terminology, the *I* quarter lag includes nonzero coefficients for the independent variable lagged zero through *I* – 1 quarters, and a zero coefficient for the independent variable lagged *I* or more quarters. In Cukierman's terms, this is an *I* – 1 quarter lag.

TABLE 1—FURTHER ESTIMATES OF THE MONEY-ILLUSION CONSUMPTION FUNCTION

Equa- tion	Independent Variables							Statistics		
	Constant	$lny$		$lnw$	$lnP$			$R^2$	$S.E. \times 10^2 D.W.$	
		$I$	$\Sigma \gamma_i$	$\delta_0$		$K$	$\Sigma \eta_k$			
1-1	-1.953 ( 0.114)	7	0.661 (0.043)	0.127 (0.036)		7	0.148 (0.036)	.9984	.2964	1.76
				$\delta_0$	$\delta_1$	$\eta_0$	$\eta_1$	$\eta_2$		
1-2	-1.947 ( 0.136)	8	0.623 (0.043)	0.177 (0.053)	-0.030 (0.060)	0.060 (0.130)	0.326 (0.149)	0.024 (0.117)	.9982	.3044 1.63
1-3	-1.992 ( 0.103)	4	0.596 (0.035)	0.156 (0.035)		0.075 (0.112)	0.362 (0.149)	-0.018 (0.113)	.9981	.3071 1.48
						$K$	$\Sigma \eta_k$			
1-4	-2.024 ( 0.102)	4	0.602 (0.035)	0.146 (0.035)		3	0.430 (0.034)		.9980	.3107 1.61

function in our original paper. (This estimate also appears as equation 1-1 in Cukierman's Table 1.) Equations 1-2 and 1-3 of Table 1 correspond to Cukierman's equations 1-2 and 1-3 with the exception that we use the aggregate *CPI* while he disaggregates the *CPI* into its five components. The income coefficients in equations 1-2 and 1-3 were estimated using the Almon lag technique with a third-degree polynomial and the last (*I*th) coefficient constrained to zero. In contrast, since the wealth and price lags in these equations are short, we have estimated the coefficients of these lags unconstrained, including the relevant current and lagged values of these independent variables directly in the equation. Our equation 1-4 simply repeats 1-3 with the price lag reestimated using the Almon technique instead of the separate appearance of  $P_t$ ,  $P_{t-1}$ , and  $P_{t-2}$ .

Comparing our results with Cukierman's, one sees that in equation 1-2 disaggregation of the price variable reduces the income and wealth coefficients somewhat. But disaggregation of the *CPI* leaves the sum of the price coefficients almost unchanged—our sum is 0.410 while his is 0.414. In equations 1-3, with a four-quarter income lag but only the current wealth term, the sum of our income

coefficients is quite the same as Cukierman's; his wealth coefficient is a bit smaller than ours, and the sum of his price coefficients is 0.382, compared to our estimate of 0.419.

The striking difference between Cukierman's equation and ours occurs in the respective estimates of the standard error of the sum of the price coefficients. We have calculated the standard error of our  $\Sigma \eta_k$  in equation 1-3 from the variance-covariance matrix of the estimated price coefficients, shown in Table 2. The variance of  $\Sigma \eta_k$  from Table 2 is 0.0012, and hence the standard error is 0.035. As equation 1-4 shows, when 1-3 is reestimated applying the Almon lag technique to the price lag the resulting estimate of the sum of the price coefficients is 0.430 with a standard error of 0.034, quite similar to the estimates from equation 1-3.

Thus the difference between our results and Cukierman's centers on the estimate of the standard error of the sum of the price coefficients. The two estimates of the sum of the price coefficients are quite close, but the sum of our price coefficients in 1-3 is 12.0 times its standard error, while Cukierman's *t*-ratio is only 2.85. Our experiments with the lag lengths, shown in Table 1, suggest that this difference results from disaggregation of

TABLE 2—VARIANCE-COVARIANCE MATRIX OF ESTIMATES OF  $\eta_k$  IN EQUATION 1-3

	$\eta_0$	$\eta_1$	$\eta_2$
$\eta_0$	0.0126	-0.0116	-0.0003
$\eta_1$	-0.0116	0.0221	-0.0112
$\eta_2$	-0.0003	-0.0112	0.0127

the price series, not from changes in the lag lengths. The question, then, is how should this difference be interpreted?

## II. An Interpretation of the Results

Cukierman's disaggregated equation (4) can be written as

$$(2) \quad c_t = \beta y_t^\gamma w_t^\delta \prod_{i=1}^I P_{it}^{\eta_i}$$

where the index  $i$  runs across his five sub-components of the *CPI*, and where, for the moment, lag distributions on the independent variables are ignored. The weighted product of the price terms in (2) can, in turn, be rewritten as

$$(3) \quad \left( \prod_{i=1}^I P_{it}^{\alpha_i} \right)^\eta$$

In (3), the  $\alpha_i$ 's are geometric weights for the price series with  $\sum_i \alpha_i = 1$ , and  $\eta$  is the analogue of our estimated price coefficient. Considering Cukierman's work in light of (2) and (3), what he has done is replace the *CPI*, which weights his component series arithmetically using a priori base-year weights, with a new price index that weights those components geometrically. Moreover, he has derived the  $\alpha_i$  weights for his price series as part of his consumption-function estimation so that the  $\alpha_i$ 's implicit in his results are those which give the best fit in his consumption function. Thus Cukierman's result can be interpreted as an estimate of our money-illusion parameter of 0.382, derived using a price index with geometric weights estimated simultaneously with the other parameters in the consumption function. In contrast, our results give an estimate of 0.418—not significantly different from Cukierman's 0.382—using a base-weighted index.

The question remains, why is Cukierman's estimate of the standard error of the price coefficient so much larger than ours—0.134 versus 0.035? First, he estimates an equation with twelve separate price terms among the independent variables. As Cukierman notes at several points in his comment, the large number of price coefficients leads to a fairly severe multicollinearity problem. This collinearity problem tends to reduce our confidence in his results.

Most importantly, however, by estimating the weights in his price index, the  $\alpha_i$ 's mentioned above, Cukierman may be adjusting his equation for trend factors. Each of the five price series he uses has its own trend, as does the consumption series. We suspect that Cukierman's procedure may assign weights to the several price series partially to account for the *trend* in real consumption, leaving the *variation* in the weighted price series about its trend to explain the variation in consumption about its trend. This interpretation of Cukierman's results is supported by our original experiment with a deviation-from-trend version of the money-illusion consumption function, reported as equation 3-1 in Table 3 of our original article. In that equation explaining deviations from trend in real consumption, the sum of the price coefficients was reduced to 0.300 with a standard error of 0.110. The resulting *t*-ratio of 2.73 is close to Cukierman's estimate of 2.85 for the ratio of the sum of his price coefficients to its standard error. Cukierman's equation 1-3 seems to yield a sum of price coefficients similar to our basic equation's 0.418 with a standard error similar to our detrended equation's 0.110. Thus it seems that Cukierman's price weights may be introducing a trend factor into his equation, leaving price deviations from trend to explain consumption deviations from trend.

## III. Conclusion

From the discussion of Sections I and II of this reply, it should be clear that we interpret Cukierman's results as being consistent with ours, and perhaps even mildly support-

tive of our conclusion concerning the existence of money illusion in the *U.S.* economy. One additional element of his work also supports this interpretation.

In our article, we stated that "... if one believes that real consumption depends on present and lagged values of money income with each value deflated by a misperceived price level, then one would think the price lag and income lag should be roughly the same length since the variable really moving consumption is incorrectly deflated income"

(p. 840). Since he was working with five price series, Cukierman initially set the length of his price lags at three quarters (in our terminology). He then found that the "best" income lag with that price lag was four quarters (again in our terminology) rather than our seven-quarter income lag. Thus his results, as ours, suggest that the income and price lags should be of roughly the same length, supporting the view that the variable really moving consumption is incorrectly deflated income.

# Do Blacks Save More?

By MARJORIE GALENSON\*

According to the *Study of Consumer Purchases, 1935-36*, blacks saved more than whites within comparable income classes. This surprising finding played an important role in the history of theories of the consumption function. The average for blacks of all incomes in the survey was lower than for whites, of course, because of their greater concentration at low incomes. But the results that emerged after standardizing for income seemed paradoxical, and aroused a good deal of interest.

After a brief review in Sections I and II of the literature on black-white savings differentials, based mainly on data from the government surveys of 1935-36 and 1950-51, I shall present in Section III an analysis of the *Survey of Consumer Expenditures 1960-61*, showing the effect on the comparison of the elimination of statistically abnormal families (single persons; heads over 65; income extremes) and of James Tobin's technique of stratifying by savers and dissavers in order to control for assets. The result disposes fairly conclusively of the widely held view that blacks save more.

## I. *The Study of Consumer Purchases, 1935-36*

Horst Mendershausen was the first to analyze the savings differential in print, in 1940. He found the differences in thriftiness between blacks and whites to be statistically significant. The next contribution was from Richard Sterner, in *The Negro's Share*, published in 1943: he observed that blacks had smaller net deficits than whites at low incomes and, at higher incomes, larger surpluses. He reasoned that their higher savings

were due to the fact that blacks had less access to credit, and were less likely than whites, at any given income level, to have seen better days, or to have expectations of higher incomes in the future—a remarkable intuitive anticipation of later theories, particularly Milton Friedman's.

The next developments in the consideration of the race differential are better known because they are (or at least were) taught in first-year theory courses. Dorothy Brady and Rose Friedman discovered that the disparity between the savings ratios disappeared when blacks were compared to whites at the same percentile of the income distribution, rather than at the same absolute income levels. They did not attempt an explanation. James Duesenberry, independently making the same findings, used the race differential to good effect in his "relative income" theory of consumption, which supplied a rationale for it: blacks and whites were in effect two different communities, and had different patterns of spending and saving. The black community was segregated and also poorer: "its members will make fewer unfavorable comparisons between their consumption standards and that of their associates;" therefore the upward pressure on consumption caused by emulation was weaker for them than for whites (pp. 48-50). This explanation seems thin on its face, and was further attenuated by his own observation that even in the 1930's, before pervasive television advertising, the two groups were "subjected to the same ways of doing things. Moreover, the ranking of goods is about the same in the two communities" (p. 50). Duesenberry, in short, accepted the 1935-36 figures as evidence of greater black thriftiness and concluded that the black savings-consumption pattern was simply different from whites'.

The defense of the traditional consumption function, with absolute income rather than relative income as the determinant of

\* Assistant professor, department of consumer economics and public policy, New York State College of Human Ecology, Cornell University. I am indebted to F. G. Pyatt, Walter Galenson, Alice Galenson, and the managing editor for good advice at various stages of this investigation, which was supported in part by a grant from the Welfare Administration and Social Security Administration, U.S. Department of Health, Education and Welfare, Washington, D.C.

consumption, was then undertaken by Tobin, with one important modification: he added assets to the equation. So much is generally known. However, many economists do not seem to be aware that in the same article Tobin explained away the apparently higher black savings.

Data on the asset holdings of whites and blacks were not available to Tobin,<sup>1</sup> but it was plausible to assume "that Negroes had, on the whole, smaller financial resources other than income. Consequently, Negroes were unable to dissave as frequently or as much as whites" (p. 144).

Tobin argued that there would be some savers and dissavers in *every* income class, not only in the income classes with a net average positive or negative saving, respectively; and the greater *dissaving* among whites, made possible by their superior financial reserves, would pull down the savings-income ratio for whites in every income class.

In any sample of families, white or Negro, the net average saving for every income group conceals a wide variation in saving by individual families. At almost every income level, there are some families who save and some who dissave. The proportion of savers rises and the proportion of dissavers falls as income increases.

The saving behavior of families below their break-even points will depend on the financial resources at their command. If they have no asset holdings or credit, their saving will be zero or positive to the amount of their contractual saving. Availability of credit is to a large extent dependent on asset holding; and families are, if only because of the costs of borrowing, more willing to draw on their own savings than go into debt. Therefore, the asset holdings of the families below their break-even points will determine the amount of their dissaving.

[pp. 146-48]

Tobin's charts on savings of whites and blacks in two cities, stratified into savers and dissavers—a rough method of controlling for differences in assets—gave striking support to his argument. The paradox was solved.

A separate comparison, using subsamples of the same family type (husband-and-wife families, two-parent families with one or two children under 16), to control for differences in family composition, reinforced the conclusions based on the two complete samples (p. 147).

Milton Friedman took the black-white differential analysis one step further. To Friedman the crucial datum was that the blacks had, as a group, lower mean income and savings than whites, evidence for him that, at the same measured income, the blacks had in fact lower "permanent incomes." Friedman's contribution could be considered a more succinct version of Tobin's explanation, since his consumption function included wealth in the equation along with "permanent" income.<sup>2</sup>

All these analyses of black-white differences were based in whole or in large part on the *Study of Consumer Purchases, 1935-36*. Those using the data considered them to be seriously deficient because of the exclusion from the sample of relief families, broken families, non-relief families with very low incomes, and families having more than one roomer or boarder, exclusions which could be expected to affect blacks more than whites. In fact, Friedman considered sampling bias important enough to account for the observed differences between white and black savings (pp. 80-81, fn. 36).

Laurence Klein and Horace Mooney, using Michigan Survey Research Center data for 1947-50, confirmed the "existence" of race differences in saving—but with puzzling changes of direction: it appeared, according to these data, that only in the North did blacks save more; in the South, on the con-

<sup>1</sup> Data on assets and net worth by race were first gathered in a 1967 *Survey of Economic Opportunity*. They are cited in Brimmer and Terrell. The survey revealed large differences, particularly at the lower income levels.

<sup>2</sup> Friedman himself regarded his permanent income theory as superior to the relative income hypotheses, and the latter in turn as superior to the absolute income hypothesis (pp. 169-82).

trary, whites saved more. This held when durables were included in savings, as they were here for the first time. Another interesting finding was that credit availability did not seem to constitute a greater constraint on blacks than on whites. Among families that had bought furniture or household appliances during the year, a larger proportion of blacks than whites, at comparable income levels, were using installment credit (p. 450).

### II. *The Survey of Consumer Expenditures, Income and Savings, 1950-51*

The 1950-51 *SCE* did not suffer from the sampling defects of the 1935-36 survey. It showed the same black-white savings differences within income classes. Here too, the apparent thriftiness of blacks was attributable, Irwin Friend and Stanley Schor ascertained, to the much higher *dissaving* in cash and deposits by whites (p. 230).

The evidence on the black-white savings differentials, particularly the Tobin and Friend and Schor articles, did not appear to be widely known among economists. The statement could be made a few years later in a reputable journal that "The theory

prevails in current economic literature that at any given income level Negroes save more than whites . . ." (Broadus Sawyer, p. 217).

### III. *The Survey of Consumer Expenditures, 1960-61*

Table 1 tells the usual story: for six income classes out of nine, blacks saved more than whites. There was no obvious reason in the socioeconomic characteristics of families in the sample why blacks should save more. On the contrary, black families were larger and contained more children under 18, conditions generally thought to make saving more difficult; and fewer blacks than whites owned their own homes, a major form of saving for families with average and below-average incomes.

The published data, and Table 1, included single-person households with families. Because I am interested in the behavior of normal families, I took advantage of the availability of the electronic General Purpose Tape, containing data by individual household, to eliminate from the study not only single persons but households headed by persons over 65 years, which together constituted 26 percent of the black sample

TABLE 1—SAVINGS OF ALL FAMILIES AND SINGLE CONSUMERS BY RACE AND INCOME CLASS, 1960-61

Income Class	Average Income <sup>a</sup>		Average Savings <sup>b</sup>		Savings-Income Ratios	
	White	Black	White	Black	White	Black
All incomes <sup>c</sup>	\$6,247	\$3,930	\$531	\$187	.09	.05
Under \$1,000	769	941	-841	51	-1.09	.05
\$ 1,000- 1,999	1,846	1,862	-217	-28	-.12	-.02
\$ 2,000- 2,999	2,873	2,820	-124	57	-.04	.02
\$ 3,000- 3,999	3,923	3,901	-64	71	-.02	.02
\$ 4,000- 4,999	5,051	4,863	237	173	.05	.04
\$ 5,000- 5,999	6,027	5,753	377	192	.06	.03
\$ 6,000- 7,499	7,254	6,979	593	697	.08	.10
\$ 7,500- 9,999	9,087	8,823	1,032	1,198	.11	.14
\$10,000-14,999	12,205	12,458	1,811	1,188	.15	.10

Source: *Survey of Consumer Expenditures 1960-61*, table 22a and b.

<sup>a</sup> Income: Money income after taxes, other money receipts, the value of items received without expense, and the account balancing difference (sign ignored).

<sup>b</sup> Savings: Net change in assets and liabilities plus personal insurance.

<sup>c</sup> The averages for whites include families and single consumers with incomes of \$15,000 and over. There were no blacks in this income class.

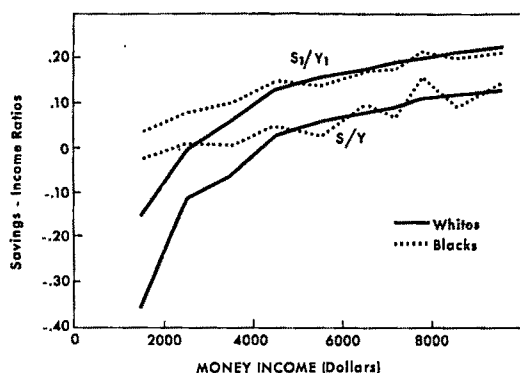


FIGURE 1. SAVINGS PATTERNS OF WHITE AND BLACK "NORMAL" FAMILIES, 1960-61.

$S$  = Net change in assets and liabilities; personal insurance

$Y$  = Money income after taxes; other money receipts  
 $S_1 = S$  plus durables

$Y_1 = Y$  plus value of items received without expense; value of home produced food; account balancing difference

and 28 percent of the white. The omission of extreme income classes—under \$1,000 and over \$10,000—reduced the whites further, to 7,571, or 62 percent of the original white sample; and the blacks to 921, or 68 percent of the original black sample.

The effect of "normalizing" the sample was unexpected. Figure 1 shows that, among families of two or more, with heads under 65, and incomes from \$1,000 to \$10,000 the whites saved, on the average, more than the blacks in three of the six income classes above \$5,000.<sup>3</sup> Below \$5,000, blacks saved when whites dissaved, or dissaved less than whites in every instance. The excluded households—particularly the single persons, the very low incomes, and the aged—must have constituted a large proportion of the dissavers, and dissavers were relatively more numerous among the whites.

An alternative pair of definitions of savings and income used in Figure 1: (a) included durables in savings, which might be expected to raise white more than black savings because of the higher proportion of white homeowners; and (b) included income in kind in the definition of income, which could be expected to raise incomes of blacks more than

whites because of their disproportionate representation in domestic service and other occupations where income in kind is customary. Raising the numerator of the savings-income ratio for the whites, and the denominator for the blacks, did indeed raise the whites' ratio above the blacks in several income classes.

Tobin's technique of stratifying the 1935-36 sample into savers and dissavers was repeated with the "normal" families in the 1960-61 survey. Figure 2 charts the data for the narrower definitions of savings and income, and Figure 3 for the more inclusive. They show very clearly that, among savers, whites saved slightly more than blacks in every income class, except for a tie in the \$7,500-8,000 class in Figure 2. It is also clear that whites engaged to a greater extent in deficit financing.

These tests confirm the significance of assets in explaining the apparent black-white savings differential, and the necessity of

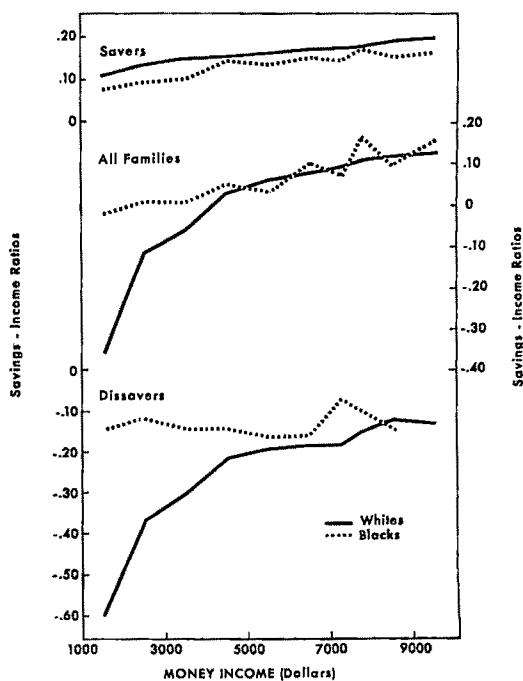


FIGURE 2. SAVINGS PATTERNS OF WHITE AND BLACK "NORMAL" FAMILIES, 1960-61.

$S$  = Net change in assets and liabilities; personal insurance

$Y$  = Money income after taxes; other money receipts

<sup>3</sup> The tables on which the graphs are based are available on request.

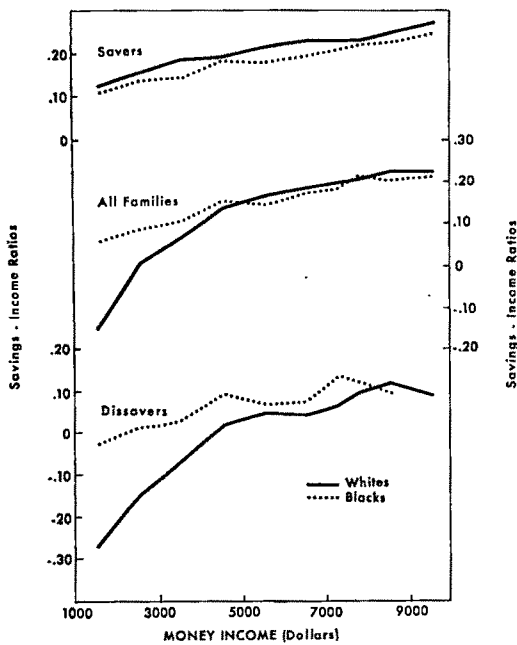


FIGURE 3. SAVINGS PATTERNS OF WHITE AND BLACK "NORMAL" FAMILIES, 1960-61.

$S_1$  = Net change in assets and liabilities; personal insurance; durables

$Y_1$  = Money income after taxes; other money receipts; value of items received without expense; value of home produced food; account balancing difference.

including them with income in the consumption function. They can also be interpreted in terms of the permanent income theory. Because the average income of white families—corresponding, by Friedman's definition, to the permanent income of the group—is higher than the figure for blacks, the observed income of whites with incomes below the average is likely to contain a larger negative transitory component than the same level of measured income of black families. At the same low income level, whites can on the average support a higher (permanent) consumption level than blacks because they have financial reserves to draw on. At higher levels, where more blacks than whites will be above their permanent income, we should expect to find a larger positive transitory element in the incomes of the blacks—which means higher savings. Figure 2 (all families) does indicate that blacks saved more in half the income classes over \$5,000, but in Figure

3, where durables are included in savings, whites saved more on the whole, although the differences were small.

#### IV. Summary

The *Survey of Consumer Expenditures 1960-61* confirms Tobin's finding, based on the 1935-36 data, and that of Friend and Schor using 1950 data: blacks do not save more; whites merely *dissave* more. This holds no matter which of several definitions of savings and income are used. A further finding is that the elimination of "abnormal" families served to control, at least in part, for differences in assets, which are, after income, the crucial element in explaining differences in savings behavior.

The notion that blacks saved more than whites gained currency on the basis of statistical evidence for which no cogent reason was ever supplied. Statistics that seem to deny common sense should be interpreted with caution.

#### REFERENCES

- D. S. Brady and R. D. Friedman, "Savings and the Income Distribution," *Nat. Bur. Econ. Res. Stud. in Income and Wealth*, Vol. 10, New York 1947, 247-65.
- A. F. Brimmer and H. S. Terrell, "The Economic Potential of Black Capitalism," 82nd Annual Meeting of the American Economic Association, New York, Dec. 29, 1969.
- J. S. Duesenberry, *Income, Saving and the Theory of Consumer Behavior*, New York 1967.
- M. Friedman, *A Theory of the Consumption Function*, Princeton 1957.
- I. Friend and S. Schor, "Who Saves?" *Rev. Econ. Statist.*, May 1959, 41, 213-45.
- L. R. Klein and H. W. Mooney, "Negro-White Savings Differentials and the Consumption Function Problem," *Econometrica*, July 1953, 21, 425-56.
- H. Mendershausen, "Differences in Family Savings Between Cities of Different Size and Location, Whites and Negroes," *Rev. Econ. Statist.*, 1940, 22, 122-37.
- B. Sawyer, "An Examination of Race as a Factor in Negro-White Consumption Patterns," *Rev. Econ. Statist.*, May 1962, 44, 217-20.
- R. Sterner, *The Negro's Share*, New York 1943.

- J. Tobin, "Relative Income, Absolute Income, and Savings," in *Money, Trade and Economic Growth*, New York 1951, 135-56.
- U.S. Bureau of Labor Statistics, *Study of Consumer Purchases, 1935-36*, Bulls. 642-49, Washington 1939-41.
- , *Study of Consumer Expenditures, Income and Savings, 1950*, 18 vols., Univ. Pennsylvania, Philadelphia 1956-57.
- , *Survey of Consumer Expenditures, 1960-61*, BLS reports, series 237, Washington 1964-66.

# On Measuring the Nearness of Near-Moneys: Comment

By TONG HUN LEE\*

In a recent paper published in this *Review*, V. K. Chetty developed an interesting method of estimating the substitution parameters between liquid assets and found that commercial bank time and savings deposits, mutual savings bank deposits, and savings and loan association shares (*T*, *MS*, and *SL* hereafter) rank in the descending order of importance as near-moneys. While Chetty suggested inclusion of these near-moneys in the definition of money, he argued that since *T* are closer substitutes for money than are *SL*, Milton Friedman's definition of money including *T* but not *SL* is also justified. Indeed, Friedman and Anna Schwartz (1970, p. 188) subsequently claimed that his definition of money is confirmed by Chetty's results. Chetty's finding, however, depends critically on the incorporation in his analysis of pre-1951 observations which do not reflect an important institutional change that occurred in 1951 affecting the substitution parameters. Omitting such observations but using the same method, this paper will show that *SL* are closer substitutes for money than are *T*. Although this result does not affect the basic methodological contribution made by Chetty, it reverses his empirical finding. While I will not argue for including *SL* in the definition of money for the reasons discussed later, the present finding has an important implication for rejecting Friedman's concept of money now widely accepted in monetary analyses.

Chetty employed 1945-66 annual time-series observations for his analysis, but the analysis should have been confined to a period after 1950. In the fall of 1950, the insurance provision of the Federal Savings

and Loan Insurance Corporation was made more liberal than before in the event of default of an insured savings and loan association. As noted by Friedman and Schwartz (1963, p. 669), this provision in fact became identical to that governing the Federal Deposit Insurance Corporation. It is, therefore, quite reasonable to expect that the nearness of *SL* to money has increased since 1951. Indeed, there is evidence (see G. K. Kardouche, Lee (1966)) showing that the substitutability of *SL* for money has shifted due to the institutional change cited here. In addition, the post-1950 data pertain to the period of revival of monetary policy since the 1951 Accord and the analysis of substitution effects in a policy context should be directed to such data.

With the 1951-66 data, Chetty's estimating equations are recomputed. Since there was evidence of autocorrelation among the least squares residuals, the equations are reestimated to remove autocorrelation by using Phoebe Dhrymes' method with the following results:<sup>1</sup>

$$(1) \quad \log T = .0180 - 38.81 \log \frac{1}{1 + r_T} \\ (.061) \quad (2.94) \\ + .6668 \log M \quad \bar{R}^2 = .983 \quad DW = 1.81 \\ (.024)$$

<sup>1</sup> Dhrymes' estimation procedure used is as follows: In each estimating equation all the variables are transformed by taking the original observation and subtracting autocorrelation coefficient times the observation lagged one period. Unlike the traditional textbook method, the first observation is also used by multiplying the square root of one minus the squared autocorrelation coefficient. By changing the value of the autocorrelation coefficient for alternative runs of the regression, the regression equation which minimizes the standard error of the equation is chosen as the estimated equation presented above. The estimated autocorrelations for *T*, *MS*, and *SL* equations are .62, .45, and .69, respectively. The estimators involved in each equation are consistent and asymptotically more efficient than the least squares estimators. While Dhrymes' method provided better results as expected, the least squares method suggested by Chetty also gave results consistent with the conclusions drawn in this paper.

\* Professor of economics, University of Wisconsin-Milwaukee, and a member of the Social Systems Research Institute at the University of Wisconsin, Madison. I am grateful to Thor Hultgren, John H. Makin, and anonymous referees for their helpful comments on the initial draft of this paper. This study is financed by a grant from the UWM Graduate School.

$$\begin{aligned} \log MS &= .0627 - 32.23 \log \frac{1}{1 + r_{MS}} \\ &\quad (.068) \quad (2.22) \\ + .4829 \log M \quad \bar{R}^2 &= .982 \quad DW = 1.72 \\ &\quad (.020) \\ \log SL &= .3060 - 72.07 \log \frac{1}{1 + r_{SL}} \\ &\quad (.084) \quad (9.49) \\ + .1102 \log M \quad \bar{R}^2 &= .971 \quad DW = 1.80 \\ &\quad (.041) \end{aligned}$$

where  $M$  denotes the sum of currency and demand deposits,  $r$ 's indicate the respective interest rates, and figures in parentheses are standard errors of the estimates. Following Chetty, the implied utility function of holding liquid assets is derived from the above equations as:

$$\begin{aligned} (2) \quad U &= [M^{.993} + 1.01T^{.974} + 1.02MS^{.969} \\ &\quad (.09) (.005) \quad (.08) (.006) \\ &\quad + 1.02SL^{.986}]^{1.007} \\ &\quad (.06) (.001) \end{aligned}$$

Asymptotic standard errors of the above parameter estimates were derived from the covariance matrix of the coefficient estimates in (1) by employing Lawrence Klein's method. Following Chetty, the exponent of  $M$  is obtained by the weighted average of three estimates of the parameter by using the reciprocals of the respective variances as their weights.<sup>2</sup> As a result, standard errors are not shown for this exponent and its reciprocal (outside the bracket).

It is important to assess the impact of the difference between Chetty's estimates and those in equation (2) upon the substitutability of various assets and money. Chetty argued that all of the exponents in the utility function were almost unity and, therefore, that relative size of the coefficients on liquid assets determined the nearness of near-moneys. Specifically he found that the coeffi-

cients on  $T$ ,  $MS$ , and  $SL$  were 1.02, .880, and .616, respectively, thereby concluding that  $T$ ,  $MS$ , and  $SL$  rank in that order as near-moneys.

Note, however, that when the pre-1951 observations are omitted from Chetty's data, the *coefficients* of utility function (2) on  $T$ ,  $MS$ , and  $SL$  are not significantly different from unity and virtually the same. On the other hand, the *exponent* of  $SL$  is greater than that of  $T$  (and also that of  $MS$ ), while the respective exponents are significantly smaller than one.<sup>3</sup> A larger value of the exponent of  $SL$ , given the virtual identity of the coefficients, measures a higher degree of substitutability of money and  $SL$ . Thus, using Chetty's criterion, the moneyness of savings and loan shares exceeds that of other liquid assets considered. This conclusion is reinforced by the relative magnitudes of the Hicks-Allen partial elasticities of substitution computed as follows:

$$\begin{aligned} (3) \quad \sigma_{M,T} &= 43.79 \quad \sigma_{M,MS} = 35.75 \\ \sigma_{M,SL} &= 95.46 \end{aligned}$$

The elasticity of substitution between  $M$  and  $SL$ ,  $\sigma_{M,SL}$ , is more than twice of  $\sigma_{M,T}$  as opposed to Chetty's estimates of 35.46 for the former and of 30.86 for the latter.

Since  $SL$  and  $MS$  are owned mostly by households while  $T$  are owned by both households and business firms, it may seem more appropriate to use data on household money holdings alone rather than total money holdings. At least household data could be

<sup>2</sup> In estimating the general CES type function in (2) by Chetty's method, one would obtain as many estimates of the exponent of  $M$  as there are estimating equations. Chetty informed me that he used the conventional method of weighting such estimates by using the reciprocals of the respective variances.

<sup>3</sup> Statistical significance referred to in this and the preceding sentences are based on the large sample test using asymptotic standard errors and .01 significance level. Also if one assumes that the disturbance terms of the estimating equations are mutually independent and normally distributed, one could test, on the basis of large sample properties, the null hypothesis that the *exponent* of  $SL$  is equal to that of  $T$ . This hypothesis was rejected at .01 significance level in favor of the alternative hypothesis that the exponent of  $SL$  was greater than that of  $T$ . On the other hand, a similar test showed that the *coefficients* on  $SL$  and  $T$  were not significantly different at .01 level. Although these two tests were not carried out jointly and are not exact for a finite sample, such approximate tests are of course better than no tests at all.

used to check the accuracy of the aggregate data for measuring the substitutability of money and liquid assets. Calculations with household data from the Flow of Funds Accounts, however, clearly confirmed results from aggregate data.<sup>4</sup>

The analysis, therefore, shows that since 1951 the nearness of *SL* to money is greater than that of *T*. This result is consistent with the findings of my earlier studies (1966, 1967, 1969) in which using a different method I employed a variety of data, namely, both annual and quarterly time-series data, temporal household cross section data, and temporal aggregate cross section data by states.<sup>5</sup>

A reason for obtaining a lower degree of substitution of *T* for money relative to that

<sup>4</sup> Since the Flow of Funds Accounts do not report separate household data for *MS* and *SL*, the sum of *MS* and *SL* was utilized for estimating the utility function and the elasticities of substitution. The corresponding interest rate used was the average of rates on *MS* and *SL* weighted by the respective deposit sizes. The results so computed for the 1951-66 period are as follows:

$$U = [M^{.998} + 1.04 T^{.971} + 1.04 (MS + SL)^{.986}]^{1.002} \\ (.04) (.001) (.03) (.001)$$

$$\sigma_{M,T} = 43.08$$

$$\sigma_{M,(MS+SL)} = 142.30$$

While there are some differences for obvious reasons, the above results clearly confirm those obtained from aggregate data. Since  $\sigma_{M,(MS+SL)}$  is greater than  $\sigma_{M,T}$ , it is very likely that  $\sigma_{M,SL}$  is also greater than  $\sigma_{M,T}$  for consumers. Moreover, the relative size of  $\sigma_{M,(MS+SL)}$  to  $\sigma_{M,T}$  in the case of households is much larger than the relative size of  $\sigma_{M,SL}$  or  $\sigma_{M,MS}$  to  $\sigma_{M,T}$  in the case of both households and business firms. In other words, the substitution hypothesis of non-bank intermediary liabilities that is primarily based on household behavior has a stronger support from household data. I am indebted to an anonymous referee for his helpful comments which led me to prepare this footnote and related paragraph in the main text.

<sup>5</sup> Mistakenly, Chetty assigns to me the incorrect view that *T* are not substitutes for money. In my previous work I clearly stated the opposite with empirical evidence (1967, p. 1171 and fn. 8). Moreover, I wrote specifically "while admitting the possible influence of a whole spectrum of interest rates on the demand for money, the principal question of the paper was: Which interest rate . . . exerts the most significant influence on the demand for money?" (1969, p. 417). Therefore, nowhere did I argue that *T* (or *MS* for that matter) are not substitutes for money.

of *SL* stems mainly from the fact that for most of the time period under study, the yield on *T* had been kept substantially lower than other yields through maximum rate control under Regulation Q. Such a regulation for governing *T* on the one hand and a liberalized insurance provision for *SL* on the other would have raised the substitutability of *SL* for money over that of *T* for this period until around mid-1960. As noted by James Tobin, if the liabilities of thrift institutions like *SL* are better substitutes for money than are *T*, Friedman's definition of money is inadequate, particularly because the necessary condition postulated by Friedman is contradicted.

I do not, however, suggest that *SL* be included in the definition of money. The inclusion of *SL* and/or even some other assets depends critically on the stability of the estimated substitution parameters if such estimates are to be utilized for the purpose of predicting money totals. Increased use of *CD*'s along with liberalized Regulation Q in recent years and also initiation of dividend ceilings on *SL* by the Federal Home Loan Bank Board since late 1966 may ultimately shift the substitution relationships among savings deposits.<sup>6</sup> Moreover, it should be noted that the existence of strong substitution effects among liquid assets are the necessary but not the sufficient condition for combining such assets for the definition of money.

## REFERENCES

- V. K. Chetty, "On Measuring the Nearness of Near-Moneys," *Amer. Econ. Rev.*, June 1969, 59, 270-81.
- P. J. Dhrymes, "On the Treatment of Certain Recurrent Non-Linearities in Regression Analysis," *Southern Econ. J.* Oct. 1966, 33, 189-96.
- M. Friedman and D. Meiselman, "The Rela-

<sup>6</sup> Instability of the estimated parameters may also be caused by a misspecified model. The CES-type function, assuming constant ratios of elasticities of substitution, may not accurately depict the utility function conceived by liquid asset holders.

- tive Stability of Monetary Velocity and the Investment Multiplier in the United States, 1897-1958," in E. C. Brown et al., eds., *Stabilization Policies, Commission on Money and Credit*, Englewood Cliffs 1963.
- M. Friedman and A. J. Schwartz, *A Monetary History of the United States 1867-1960*, Princeton 1963.
- and ———, *Monetary Statistics of the United States*, New York 1970.
- G. K. Kardouche, *The Competition for Savings*, New York 1969.
- L. R. Klein, *Textbook of Econometrics*, Evanston 1953.
- T. H. Lee, "Alternative Interest Rates and the Demand for Money: The Empirical Evidence," *Amer. Econ. Rev.*, Dec. 1967, 57, 1168-81.
- , "Alternative Interest Rates and the Demand for Money: Reply," *Amer. Econ. Rev.*, June 1969, 59, 412-18.
- , "Substitutability of Non-Bank Intermediary Liabilities for Money: The Empirical Evidence," *J. Finance*, Sept. 1966, 21, 441-57.
- J. Tobin, "The Monetary Interpretation of History," (A Review Article), *Amer. Econ. Rev.*, June 1965, 55, 464-85.

# On Measuring the Nearness of Near-Moneys: Comment

By LARRY STEINHAUER AND JOHN CHANG\*

This paper is a review of an attempt by V. K. Chetty to measure the relative amount of monetary services rendered by near-money assets such as time deposits at commercial banks, deposits at mutual savings banks, and savings and loan association shares. If such a measure could be found, then one could use it to construct a better money supply total; that is, one which would more accurately measure the total amount of monetary services available in the economy at any point of time. For example, if it could be established that a dollar of savings and loan shares ( $SL$ ) rendered the same amount of monetary services as fifty cents of money ( $M$ ) then by adding 50 percent of the value of outstanding  $SL$  to the conventional money supply, we would arrive at an adjusted money supply ( $M_a$ ) which would take into account the monetary services of  $SL$  as well as  $M$ . The  $M_a$  would then measure the amount of  $M$  alone it would take to provide the same monetary services as the actual combination of  $M$  and  $SL$  in existence.

Interest in constructing an  $M_a$  along these lines has been expressed by several writers.<sup>1</sup> Their interest stems from the belief that there exists a stable relationship between the level of income, properly defined, and the desired level of monetary services. If this is the case, then a more accurate measure of the total amount of monetary services available in the economy would enable one to more accurately predict the level of income.

Chetty's attempt to construct an  $M_a$  will be reviewed in two sections. In Section I, the theory that Chetty uses to construct  $M_a$  is examined, and it is shown that Chetty's theoretical presentation contains an error. Moreover, even if the theory is corrected

along the lines suggested by Chetty in a footnote, it would still be impossible to estimate the monetary services of near-money assets since these cannot be separated from the other nonmonetary services rendered by these assets. In Section II, Chetty's empirical results are examined. Because Chetty cannot separate the monetary from the nonmonetary services of near-money assets, he is led to assign weights to near-money assets which are greater than one, implying that these assets yield a larger amount of monetary services per dollar than does money itself.

## I

Chetty's method of assigning weights to near-money assets is rather innovative. He suggests that it might be easier to directly estimate an  $M_a$  which incorporates near-money assets and then work backwards to assign weights to these assets. Briefly, his technique is to estimate the parameters of an indifference map between money and other liquid assets and then use this function to calculate the amount of money that it would take to compensate consumers for the loss of the monetary services of near-money assets.

The mathematical expression that Chetty adopts for his utility function is similar in form to a CES production function. Such a utility function between money ( $M$ ) and time deposits at commercial banks ( $T$ ) is reproduced as equation (1) below.

$$(1) \quad U = (M^{-\rho} + \beta_2 T^{-\rho})^{-(1/\rho)}$$

In order to estimate the parameters of this utility function, Chetty must next specify the slope of the budget constraint. Chetty tries to incorporate the rate of interest in this slope. To do so he argues as follows: Suppose the consumer at the beginning of this period has liquid asset holdings of  $L_0$  dollars which consists of  $M_0$  of  $M$  and  $T_0$  of  $T$ . Then if  $T_1$

\* Assistant professors, Florida State University.

<sup>1</sup> See James Ford and T. Stark, pp. 1-3; Milton Friedman and David Meiselman, p. 185; Friedman and Anna Schwartz, pp. 2, 151-52; John Gurley, pp. 7-8; and Edward Kane, pp. 222-43.

represents the dollar value of  $T$  at the beginning of the *next* period, and if  $i$  is the rate of interest on time deposits of the current period, the budget constraint of the consumer can either be written as:

$$(2a) \quad L_0 = M_0 + T_0$$

or

$$(2b) \quad L_0 = M_0 + T_1/(1+i)$$

The slope of the budget line in which  $T_1$  is used is:  $-(1+i)$ .

The interest rate, however, cannot be incorporated in the slope of the budget constraint in this manner. For it is not necessarily true that  $T_1/(1+i)$  will be equal to  $T_0$ , if by  $T_1$  Chetty means the observed value of  $T$  at the beginning of the next period. This is simply due to the fact that the amount of  $T$  held this period does not constrain the amount that can be held next period. If a consumer desires an amount of  $T$  next period which is larger than  $T_0(1+i)$  then he will simply divert some more of his current income to the purchase of additional  $T$ . If, conversely, he desires a smaller amount of  $T$  next period than  $T_0(1+i)$ , then he will simply reduce the amount of  $T$  that he holds to the desired amount.

The proper budget constraint is the one given in equation (2a). This constraint has a slope of  $-1$ , which is equal to the price (one dollar) of a unit of  $M$  divided by the price (also one dollar) of a unit of  $T$ . The rate of interest does not appear at all.<sup>2</sup>

Since the interest rate is an important factor in any decision concerning the relative amounts of  $M$  and  $T$  to hold, the question immediately arises as to where the interest on  $T$  enters the analysis. The answer is that the proper place to take account of the interest rate is in the utility function, not the budget constraint. The interest payments on a dollar of  $T$  represent the claim to a gen-

eralized stream of future consumption services. The consumer who chooses to hold a dollar of  $T$ , therefore, receives utility not only from the direct stream of monetary services of  $T$  but from the generalized stream of consumption services as well. Since the interest payments on  $T$  are a potential source of future utility to consumers, any utility function must take these payments into account along with the utility derived from the future monetary services of  $T$ .

Chetty, in a footnote, suggests this approach himself. He claims that his results would be changed in no essential way if a utility function is adopted which explicitly takes the interest rate into account. The utility function that Chetty apparently has in mind is one of the form:

$$(3) \quad U = [M^{-\rho} + \beta_2(1+i)T^{-\rho}]^{-1/\rho}$$

If this function is substituted for the one which appears in Chetty's paper then Chetty is correct in asserting that his results will not change. The parameters estimated for this new utility function will be exactly the same as those estimated for the original function and the estimated  $M_a$  will be the same as well.<sup>3</sup>

<sup>3</sup> We can demonstrate that the estimated parameters of both utility functions will be the same as follows. The slope of a representative indifference curve from the new utility function is:

$$\partial T / \partial M = \frac{1}{-\beta_2(1+i)} \left( \frac{M}{T} \right)^{-\rho-1}$$

If we then equate the slope of the indifference curve to the slope of the budget line in order to maximize consumer utility we have:

$$1 = \frac{1}{\beta_2(1+i)} \left( \frac{M}{T} \right)^{-\rho-1}$$

But this is exactly the same equation that Chetty ends up with by equating the slope of the utility function to the slope of the budget constraint in his paper. Therefore, the estimates of the parameters  $\rho$  and  $\beta_2$  derived from this equation will be the same for either utility function. That the estimates of the  $M_a$  will also be the same is easy enough to prove. Calculating Chetty's  $M_a$  involves finding the amount of  $M$  alone which would yield the same monetary services as the actual combination of  $M$  and  $T$  held. To do this, we must first separate the monetary services of  $T$  from its generalized stream of consumption services. This can be done by simply assuming that the interest rate is equal to zero so that  $T$

<sup>2</sup> Once the slope of the budget line is conceived as the ratio of the price of  $M$  to the price of  $T$ , it is apparent why the rate of interest does not enter. The price of  $T$  has the dimension of dollars. The interest rate on  $T$ , on the other hand, is measured in dollars per dollar per year. The two, therefore, cannot be added together since they are dimensionally different quantities.

However, this does not mean that the analysis is correct. For Chetty's utility function still contains a hidden assumption that invalidates his technique for constructing an  $M_a$ . In order to reveal this assumption, suppose for the following discussion that the rate of interest on  $T$  is equal to zero. If this is the case, then the only utility that the consumer will derive from  $T$  is that which he receives from the monetary services of  $T$ . Furthermore, since the monetary services of  $T$  must always be less than those of  $M$  the slope of any indifference curve between  $M$  and  $T$ ,  $\partial T/\partial M$ , must always be greater than one. This is because it will always take more than one dollar of  $T$  to compensate for the foregone monetary services of a dollar of  $M$ . However, if we look at the slope of a representative indifference curve from Chetty's utility function, given in equation (4), it is clear that if the value of  $M/T$  is large enough, the slope of the indifference curve will be less than one,<sup>4</sup> where

$$(4) \quad \partial T/\partial M = -\frac{1}{\beta_2} \left( \frac{M}{T} \right)^{-\rho-1}$$

The only way to rationalize the shape that Chetty gives to his indifference curves is to assume that, in addition to its monetary services,  $T$  provides other nonmonetary services as well. Thus, as we move along the indifference curve between  $M$  and  $T$ , consumers must be compensated with enough  $M$  to make up for the loss not only of the monetary services of  $T$  but also for any nonmonetary services as well. If, for example, the marginal dollar of  $M$  yields monetary services valued at the rate of \$ .04 per year

yields no generalized consumption services. The indifference curve running through the actual combination of  $M$  and  $T$  can then be used to calculate the amount of  $M$  alone which would provide an equal value of monetary services. However, if the interest rate is set equal to zero then the new utility function reduces to the old one. Therefore, the estimated  $M_a$  using either function will be the same.

<sup>4</sup> The conditions under which the slope of a representative indifference curve will be less than one are:

$$\partial T/\partial M = \frac{1}{\beta_2} \left( \frac{M}{T} \right)^{-\rho-1} < 1 \quad \text{if } T < \beta_2^{(\rho+1)} M$$

while the marginal dollar of  $T$  yields monetary services valued at \$ .02 and nonmonetary services valued at \$ .04 per year, then even though the monetary services of  $M$  are more valuable than those of  $T$ , it would still take \$1.50 of  $M$  to compensate for the foregone total services of the marginal dollar of  $T$ .

Once the possibility of  $T$  rendering direct nonmonetary services is introduced, however, Chetty's technique for estimating an  $M_a$  is no longer valid. As explained in footnote 3, Chetty's  $M_a$  is calculated by setting the interest rate on  $T$  equal to zero and then using the indifference curve that runs through the actual combination of  $M$  and  $T$  to find the amount of  $M$  alone which would provide an equal amount of satisfaction to the consumer. However, in substituting  $M$  for  $T$  along the indifference curve, Chetty must substitute enough  $M$  to make up not only for the monetary services of  $T$  but also for the nonmonetary services as well. Thus, the  $M_a$  that Chetty calculates overestimates the true amount of monetary services contributed by  $T$ .<sup>5</sup>

The only way to correct Chetty's analysis is to find some method of separating the monetary services of  $T$  from its nonmonetary services. But there is no apparent way of doing so given the form of Chetty's utility function. Therefore, Chetty's analysis cannot be readily modified to overcome this problem.

## II

The degree of overestimation introduced into Chetty's analysis by his inability to separate the monetary services of near-money assets from their nonmonetary services becomes apparent when Chetty's empirical work is reviewed.

Chetty's calculations for  $M_a$  using  $M$  and

<sup>5</sup> For example, if as before the marginal dollar of  $M$  rendered monetary services valued at \$.04, while the marginal dollar of  $T$  rendered \$.02 of monetary services and \$.04 worth of nonmonetary services then \$.50 should be added to the  $M_a$  to compensate for the loss of the monetary services of the marginal dollar of  $T$  yet Chetty's technique would add \$1.50. This would imply, erroneously, that the marginal dollar of  $T$  rendered a larger amount of monetary services than  $M$  itself.

TABLE 1—ADJUSTED MONEY STOCK BASED ON  
MONEY AND TIME DEPOSITS AT  
COMMERCIAL BANKS ( $T$ )

Year	$M+T$	$M_a$	Average Moneyness of $T$
1945	132.4	133.5	1.037
1946	143.8	145.0	1.036
1947	148.8	150.1	1.037
1948	147.4	148.6	1.034
1949	147.3	148.5	1.033
1950	154.0	155.3	1.036
1951	162.0	163.8	1.048
1952	169.7	171.2	1.037
1953	174.2	175.8	1.037
1954	181.2	182.9	1.036
1955	186.6	188.4	1.037
1956	190.3	192.4	1.042
1957	194.7	196.9	1.039
1958	207.4	109.8	1.038
1959	212.2	214.8	1.039
1960	216.8	219.7	1.040
1961	212.2	214.4	1.035
1962	248.3	252.1	1.039
1963	268.3	272.6	1.039
1964	289.2	294.0	1.038
1965	317.2	322.8	1.039
1966	335.4	341.9	1.041

time deposits at commercial banks are reproduced in Table 1. Notice that the  $M_a$  is *larger* than the sum of  $M$  plus time deposits. If we were to take these results seriously, we would have to conclude that time deposits yield a larger amount of monetary services on average than  $M$  itself.

Chetty also estimates  $M_a$  for  $M$  and three other liquid assets taken all at once. To do this he uses a modified utility function of the form:

$$(5) \quad U = (M^{-\rho} + \beta_1 X_1^{-\rho_1} + \beta_2 X_2^{-\rho_2} + \beta_3 X_3^{-\rho_3} + \beta_4 X_4^{-\rho_4})^{-1/\rho}$$

Even though the empirical results that Chetty reports for this analysis are more plausible than the ones reported for money and time deposits alone, they are still subject to the same objection, viz., that the weights assigned to the other liquid assets will de-

pend on the monetary and nonmonetary services that they are assumed to yield, not just the monetary services. Moreover, there is an additional problem with this analysis which was not encountered in the two-asset case. This stems from the way in which Chetty is forced to estimate the parameters of the new utility function. In the course of estimation he has to estimate the value of  $\rho$  three separate times. The values that he comes up with are .985, .950, and .933. Chetty only calculates  $M_a$  using the value of  $\rho$  of .985 which turns out to be the value of  $\rho$  which is most flattering to his theory since by using this value the  $M_a$  is less than the sum of its four components. However, Chetty has no theoretical justification for selecting this value of  $\rho$  over the other two. If  $M_a$  is recomputed using the lowest es-

TABLE 2—ADJUSTED MONEY STOCK BASED ON  
MONEY, TIME DEPOSITS AT COMMERCIAL  
BANKS ( $T$ ), SAVINGS AND LOAN  
ASSOCIATIONS SHARES ( $SL$ ), AND  
DEPOSITS AT MUTUAL SAVINGS  
BANKS ( $MS$ )

Year	$M+T+MS+SL$	$M_a$ for $\rho = .985$	$M_a$ for $\rho = .933$
1945	154.9	144.75	168.6
1946	168.7	157.67	184.1
1947	175.6	164.16	191.5
1948	175.7	164.38	191.6
1949	177.8	160.20	193.7
1950	186.6	173.91	202.7
1951	197.8	183.96	214.6
1952	209.8	194.69	227.7
1953	219.4	203.33	238.3
1954	232.5	215.01	252.4
1955	244.5	225.02	264.9
1956	255.0	233.78	276.1
1957	265.6	243.53	288.4
1958	285.8	262.37	311.1
1959	297.9	272.71	324.0
1960	309.9	284.23	337.0
1961	335.3	306.57	366.2
1962	364.1	333.52	400.6
1963	396.1	364.39	437.4
1964	431.3	396.45	478.6
1965	469.0	433.73	524.4
1966	491.1	457.43	552.2

timated value of  $\rho$  instead of the highest,  $M_a$  turns out to be *greater* than the sum of its four components. On the basis of this  $M_a$ , the erroneous conclusion would again have to be drawn that at least some of the near-money assets used in constructing  $M_a$  yield a larger amount of monetary services on average than  $M$  itself. The calculations of  $M_a$  for the four-asset case appears in Table 2. It is also interesting to note that a small change in  $\rho$ , approximately 5 percent, increased the estimated  $M_a$  by as much as 20 percent. This means that a small percentage error in estimating  $\rho$  will introduce a much larger error into the estimate of  $M_a$ .

### III

The conclusion that can be drawn from the above analysis is that Chetty has not succeeded in his attempt to construct an adjusted money supply in which some measure of the moneyiness of near-money assets is included. This failure is chiefly due to the mathematical specification which Chetty gives to his indifference map. Embedded within is the assumption that near-money assets yield both monetary and nonmonetary services. Since Chetty's procedure for constructing an adjusted money supply never measures the monetary services of near-money assets separately from the non-monetary services, it assigns too great a weight to near-money assets. In fact, sometimes the procedure ends up assigning heavier weights to near-money assets than it does to money itself.

In spite of the failure to successfully construct an adjusted money supply, however,

Chetty's procedure is still an original and potentially useful one. This potential might be realized if a different mathematical specification for the utility function between money and other liquid assets could be found; one which would drop the assumption that near-money assets render nonmonetary services. To accomplish this, such a utility function would have to restrict the value of the marginal services of near-money assets so that they were always less than those of money. Finding such a function is no mean task, however, and the work on it remains to be done.

### REFERENCES

- V. K. Chetty, "On Measuring the Nearness of Near-Moneys," *Amer. Econ. Rev.*, June 1969, 59, 270-81.
- J. L. Ford and T. Stark, *Long and Short Term Interest Rates*, New York 1967.
- M. Friedman and D. Meiselman, "The Relative Stability of Monetary Velocity and the Investment Multiplier in the United States, 1897-1958," in E. C. Brown et al., eds., *Stabilization Policies, Commission on Money and Credit*, Englewood Cliffs 1963.
- M. Friedman and A. J. Schwartz, *Monetary Statistics of the United States*, New York 1970.
- J. G. Gurley, "Liquidity and Financial Institutions in the Postwar Economy," U.S. Congress, Joint Economic Committee, *Study of Employment, Growth and Price Levels*, Study Paper No. 14, Washington 1960.
- E. J. Kane, "Money as a Weighted Aggregate," *Zeitschrift für Nationalökonomie*, Sept. 1964, 3, 222-43.

# On Measuring the Nearness of Near-Moneys: Reply

By V. K. CHETTY\*

Larry Steinhauer and John Chang and Tong Hun Lee have made some critical comments of my method of aggregating various financial assets. I shall reply, first to the criticisms of Steinhauer and Chang (hereafter referred to as SC).

The first criticism of SC is concerned with the unit of measurement of time deposits. That the height of Mount Everest is 29028 feet is a meaningful and correct statement of fact even to a person who measures heights in centimeters, as long as he knows that one foot equals 30.48 centimeters; it is, of course, erroneous to any one who has not heard of the English system.

I used the value of time deposits as of the *next* period in the analysis. SC insist that only the value of time deposits in the initial period should be used. Since this involves only a change of scale of a variable, no new theorems or implications can be derived by doing this. Anticipating some misunderstanding of this point, I added a footnote to this effect in my paper. But that obviously has not been helpful, at least to SC. Hence, they remark that my theory contains an 'error.' (See fn. 2 for further discussion of this point.)

The discussion about the equality of the observed and planned values of  $T$  is irrelevant, due to the assumed absence of uncertainty.

The second criticism is that there is a hidden assumption in my paper, namely, that  $T$  provides other nonmonetary services as well. I explicitly assumed (p. 272), " $M$  and  $T$  may have some common characteristics." The way in which they reveal this "hidden assumption" is interesting. They assume that the rate of interest on  $T$  is zero. Hence, the ratio of prices is unity. Then, they show that the absolute value<sup>1</sup> of  $dT/dM$

could be less than one. Do we not know from elementary price theory that, in equilibrium, the marginal rate of substitution between  $M$  and  $T$  is equal in magnitude to the ratio of prices, which is one by assumption?

SC also insist that  $-dT/dM$  should be greater than one at all points on an indifference curve. They write, "this is because it will always take more than one dollar of  $T$  to compensate for the foregone monetary services of a dollar of  $M$ ." That the rate of interest should always be positive is only an assertion of SC. To quote Joan Robinson, "... In Adam Smith's forest, ... suppose that some hunters wish to consume more than their kill, and others wish to carry consuming power into future. Then the latter could lend to the former today, out of today's catch, against a promise of repayment in the future. The rate of interest would settle at the level which equated supply and demand for loans. Whether it was positive or negative would depend upon whether spendthrifts or prudent family men happened to predominate in the community" (p. 87).

SC's assertion that  $M_a$  is calculated by setting interest rate on  $T$  equal to zero is also wrong. They remark at several places that  $M_a$  will not truly estimate the monetary services of  $T$ . My measure *does* give the amount  $M$  needed to compensate for any  $T$  foregone so that the consumer is in the same indifference curve. Nowhere in my paper did I say that this amount of  $M$  is also equal to the magnitudes of the monetary characteristics of  $T$ . When  $M$  and  $T$  have a different number of characteristics, the task of determining the exact amount of the common characteristics (like monetary services) is ill-formulated, due to the absence of objective data on characteristics. In this context, it is also easy to see why the weight for  $T$  can be greater than one.

\* I wish to acknowledge James J. Heckman for his comments on an earlier draft.

<sup>1</sup> Throughout their note, SC write, "the slope of a representative indifference curve ... is  $\partial T/\partial M$ ." The

persistent use of the partial derivative and the phrase "the slope of a curve" puzzle me.

SC finally fault me on the use of one estimate of  $\rho$ , when I had three to choose from, and they are, again, wrong. In my paper, I did not explain how I chose the point estimate of  $\rho$ . Assuming that the estimates are independently distributed, I took a weighted average of the estimates, the weights being inversely proportional to their variances. (For the property of such estimates, see E. Malinvaud, p. 285. For a still better method, see my 1964 article.) The exponent 1.026 in the equation for  $M_a$  implies  $\rho = .974$  and not .985. Accordingly, the exponent of  $M$  should read as .974 and not .954.

In conclusion, they suggest that a theory which assumes that near-money assets do *not* render nonmonetary services will be useful. Is unrealism of a theory a virtue?

Lee has presented some estimates of the elasticity of substitution between money and various liquid assets, after omitting a few observations from the time-series data I used. It is interesting to find that estimates from the flow of funds data are almost the same as my estimates. His point estimates in general are slightly different from mine; but, Lee claims that they differ significantly. He then concludes that the savings and loan association shares ( $SL$ ) are better substitutes than time deposits ( $T$ ). However, I shall show that the tests used by Lee are not applicable to his estimates. Hence, he has not demonstrated that  $SL$  is better than  $T$ . Later on I shall also argue that the question of which asset is the best substitute for money does not have a meaningful and unique answer when there are many liquid assets.

Lee's main point is that the empirical evidence supporting the Friedman and Schwartz definition of money including time deposits, is critically dependent upon the time period I used. Lee also states that Friedman has subsequently claimed that his definition of money is confirmed by my findings. From his regressions and significance tests, Lee then concludes that Friedman's definition of money should be rejected.

First, it is worth pointing out that Friedman does not accept my findings without any reservation. In fact, he points out:

... These results confirm our own, both

in indicating that total commercial bank deposits should be used in the money aggregate rather than only demand deposits and that, for the post-war period, a still broader aggregate may be better yet.

However, while we believe that this approach is extremely promising, we have serious reservations about how much confidence can be placed in Chetty's specific results for two reasons. First, they are derived entirely from post-World War II data, which are dominated by trends. Second, on a purely theoretical level, we believe that his formulation has the defect that it makes the results depend upon a strictly arbitrary choice of the time unit used in stating interest rates. [p. 188]

Friedman and Schwartz's statement that the estimated elasticity of substitution is dependent upon the strictly arbitrary choice of the time unit, is, however, incorrect.<sup>2</sup>

<sup>2</sup> Friedman pointed out the same defect in a personal communication when he commented on an earlier version of the paper. At that time, I accepted his criticism and suggested a way to get around the problem. In retrospect, I think his criticism, and hence my solution to a nonexistent problem, are incorrect. The regression model I used was

$$\log M/T = a - \sigma \log (1 + i),$$

where  $T$  is the cash value of time deposits in the *next* period. Let the period of analysis be one year. If the interest rate is expressed in different time units, say, a month, the monthly interest rate is  $(1+i)^{1/12}$ . Since  $T$  in the left-hand side is the value of time deposits at the beginning of next year, the relative price on the right-hand side should relate to the same time period. Hence, it is  $[(1+i)^{1/12}]^{12} = 1+i$ . If one used  $(1+i)^{1/12}$ , the regression estimate of  $\sigma$  would be 12 times larger. On the other hand, one may wonder whether the same elasticity of substitution will be obtained when *both*  $T$  and the interest rates are expressed in monthly time units.

To show that this is the case, let us rewrite the first-order condition as

$$\log M - \log T_0 - \log (1 + i) = a - \sigma \log (1 + i)$$

or

$$\log M - \log T_0 = a - (\sigma - 1) \log (1 + i),$$

where  $T_0 (= T/1+i)$  is now expressed in current dollars. For the monthly regression model, we have

$$\begin{aligned} \log M - \log T_0 - 1/12 \log (1 + i) \\ = a - \sigma/12 \log (1 + i) \end{aligned}$$

or

$$\log M - \log T_0 = a - (\sigma - 1) [1/12 \log (1 + i)],$$

(over)

Lee also incorrectly attributes to me the assertion that Friedman's definition of money including time deposits and excluding  $SL$  is justified. In fact I wrote: "The correlation coefficient between  $M_a$  and  $M+T$  is .999. This does not mean that one definition is as good as the other for all purposes . . . , for purposes of controlling or explaining money supply or demand, the two series will have different implications. But if one is using money stock to predict some other variable, say, national income, then  $M+T$  is as good as  $M_a$ , or for that matter as  $M$ " (p. 479).

I fully agree with Lee that the pre-1950 observations should be omitted due to the institutional changes. Since the effect of the institutional change may be distributed over time, perhaps one should even omit a few observations after 1951. When I experimented with various time periods before I wrote my paper, the results appeared to me more or less the same. Of course, I have not performed significance tests, since it is not clear to me what these asymptotic tests mean, after repeated use of the data. Hence, I relied mainly on the point estimates and economic theory to make some plausible inferences.

Lee has used the Aspin-Welch test to find out whether the exponents of the utility function differ. The Aspin and Welch method is useful to test for significant difference of means of independently distributed random variables with different variances. The exponent of  $M$  in Lee's case is a weighted average of the estimates from all

so that the regression coefficient is still  $(\sigma-1)$ . It should be noted, however, that, in demonstrating the invariance of  $\sigma$ , we are assuming that the same model (with the same  $\sigma$ ) is true for all periods of time. This assumption is necessary, to answer questions relating to multiple periods, since my model is essentially static. Alternatively, if we assume that all periods are identical and that the flow of income from the given wealth is constant throughout the lifetime and that the consumer leaves his wealth unaltered (i.e., he consumes his interest), then it is easy to see that the problem collapses to one of making an optimal decision for the initial period. The same relation then holds at each instant of time. In this case, the value of  $T$  as of any future time will be so defined as to make the estimated elasticity of substitution independent of the time period used in the analysis.

three equations. The asymptotic variance of this estimate is also dependent upon the estimated error variances of the three equations. Hence, by construction, neither the point estimates nor their asymptotic variances are independently distributed. I do not see how Lee can *assume* independence of these estimates. If he is seriously interested in such tests, he should have tested for significant differences in estimated coefficients due to the omission of the pre-1950 observations. Lee is perhaps right in saying that in general some tests are better than no tests at all. But for the above problem, if one performs the sequence of the required tests, the level of significance could easily increase from 5 to 50 percent, for ten tests, as indicated by the Bonferroni inequality.

Even if an ingenious statistician comes up with a precise test for our problem, it is not clear to me how we could say that one asset is a better asset than another, simply because the exponent of one is statistically larger than another. The estimates of the parameters of a production function are useful, for example, to determine the output for given services of the factors of production. But the question of which of these factors is a better substitute for, say, capital, does not, in general, have one answer, especially when the degree of substitutability among factors by any definition varies with the levels of the factors and output, as is the case in our problem.

In my paper, I computed the direct partial elasticity of substitution (i.e., holding factors constant and changing one price). There are precise mathematical relations between these elasticities of substitution and the conventional cross-elasticities of substitution  $E_{ij} = (\partial \log x_i) / (\partial \log p_j)$ .<sup>3</sup> For example,  $E_{ij} = s_j \sigma_{ij}$ , where  $\sigma_{ij}$  is the Allen partial elasticity of substitution between factors  $i$  and  $j$ , holding utility constant and  $s_j$  is the share of the factor whose price has changed. Thus it is possible for one asset to have a larger Allen elasticity of substitution, but a smaller cross-price elasticity, than another asset,

<sup>3</sup> See Y. Mundlak.

since there need not be any particular relationship between any two shares and the elasticities, when there are many factors.

Since the estimated elasticities of substitution in my paper are much larger than the conventional cross-price elasticities, these results have sometimes been quoted as evidence for a much greater degree of substitution. This is again not quite correct, since the Allen partial elasticity of substitution will always be larger than the cross-price elasticity. If the cross-price elasticities are positive, the direct elasticities of substitution which I estimated, will be even larger than the Allen partial elasticity of substitution. Thus, the appropriate cross-price elasticities should be calculated, using the shares before my results are compared with those of other studies. However, for the data I used, the degree of substitutability is much larger than in other studies, even after dividing by the shares.

Lee also points out that the *SL* should not be included in the definition of money, due to the possible instability of the relationships. The basic assumption in estimating the demand for money as a function of

various interest rates and income is that the underlying utility function is stable. So I do not see any reason why predictions based on my method are affected more than any other method, when there is some instability.

#### REFERENCES

- V. Chetty, "Pooling of Cross-Section and Time-Series Data," *Econometrica*, Apr. 1964, 36, 279-90.
- M. Friedman and A. Schwartz, *Monetary Statistics of the United States*, New York 1970.
- T. H. Lee, "On Measuring the Nearness of Near-Moneys: Comment," *Amer. Econ. Rev.*, Mar. 1972, 62, 217-20.
- E. Malinvaud, *Statistical Methods of Econometrics*, Chicago 1966.
- Y. Mundlak, "Elasticities of Substitution and the Theory of Derived Demand," *Rev. Econ. Stud.*, Apr. 1968, 35, 225-35.
- J. Robinson, "The Production Function and the Theory of Capital," *Rev. Econ. Stud.*, 1954, 21, 81-106.
- L. Steinhauer and J. Chang, "On Measuring the Nearness of Near-Moneys: Comment" *Amer. Econ. Rev.*, Mar. 1972, 62, 221-25.

# Lags in the Effects of Monetary Policy: Comment

By PAUL E. SMITH\*

A number of recent studies have utilized distributed lag analysis in order to examine the timing of the effects of monetary policy. The purpose of this note is to test further the hypothesis, first suggested by Donald Tucker, that the responsiveness of aggregate private demand to changes in the money supply depends upon the time path of the response of the demand for money to changes in the interest rate, as well as the response of private expenditures to changes in the interest rate. The conclusion to be drawn from Tucker's paper, insofar as monetary policy is concerned, is that aggregate demand may react rapidly to monetary policy in spite of a long distributed lag in the product market. The interest rate may over-respond to an initial change in the money supply, due to a distributed lag in the demand for money. Thus the effects of a possibly long lag in the goods and services sector is substantially offset. More recently, J. Ernest Tanner has statistically tested the Tucker hypothesis and accepted it as correct with some reservations, the major one being that monetary policy still may not be able to induce large short-run increases in demand. The required policy-induced fall in the interest rate during the first period of the policy's implementation may not be forthcoming if the interest rate is sufficiently sticky. Indeed, such large short-run fluctuations in interest rates are seldom observable in the real world.

The hypothesis to be tested here is that substantial movements in the interest rate may not be necessary if 1) there is an accelerator relationship in the investment demand equation and/or 2) monetary policy

directly affects the goods and services market via a wealth effect rather than solely operating through the money market in the traditional Keynes-Hicks-Hansen sense. The argument presented in this paper is not based so much on the notion that the following model is more "correctly" specified than Tanner's but rather that different specifications are likely to lead to substantially different conclusions.

## I. The Model

In this section the model is specified, and its parameters are statistically estimated. The notation to be used is as follows:

- $Y$  = gross national product,
- $C$  = private consumption expenditures,
- $G$  = government expenditures plus net foreign investment,
- $r$  = yield on all corporate bonds,
- $M$  = demand deposits plus currency outside banks,
- $I$  = gross private domestic investment,
- $L$  = liquid assets, i.e.,  $M$  plus time deposits in commercial banks,
- $t$  = time.

The model consists of three structural equations, each utilizing a geometric lag distribution and two identities, i.e.,

- (1)  $\hat{Y}_t \equiv \hat{C}_t + \hat{I}_t + G_t$
- (2)  $\hat{C}_t = (1 - a)[\beta_0 + (\beta_1 - \beta_2\eta)\hat{Y}_t] + \beta_2 L_t - \beta_2 a L_{t-1} + a C_{t-1}$
- (3)  $\hat{I}_t = (1 - b)[\lambda_0 + \lambda_1(\hat{Y}_t - \hat{Y}_{t-1}) + \lambda_2 \hat{r}_t + \lambda_3 t] + b I_{t-1}$
- (4)  $\hat{M}_t^d = (1 - g)[\delta_0 + \delta_1 \hat{Y}_t + \delta_2 \hat{r}_t] + g M_{t-1}$
- (5)  $\hat{M}_t^d \equiv M_t^s$

where  $(1-a)$ ,  $(1-b)$ , and  $(1-g)$  are partial adjustment coefficients and are generally as-

\* Professor of economics, University of Missouri-Columbia. I am indebted to R. F. Gilbert and Charles Chan for computational assistance and the University of Missouri for financial support via a summer faculty research grant. The comments of David Ramsey and an anonymous referee on an earlier draft were helpful but should not be blamed for any remaining errors.

sumed to lie between zero and one;  $\eta$  is the desired ratio of liquid assets to permanent income;  $\beta$  is the long-run propensity to consume from permanent income; and the hats label the endogenous variables. The model differs from that of Tanner in that liquid assets are included in the consumption function and the investment demand equation includes a linear accelerator. In addition, the private sector is disaggregated under the assumption that the lag structure for consumption demand is apt to differ from that of investment demand. A final difference is that Tanner utilizes the negative binomial distribution as opposed to the geometric lag distribution.<sup>1</sup>

Equation (2) is a consumption function first suggested by Arnold Zellner, David Huang, and L. C. Chau, and has not been modified here. The investment demand equation is standard specification aside from the trend variable, which was included in order that the estimated interest rate coefficient be negative under the presumption that, while it is not impossible for this *IS* curve to be positively sloped, such a result is highly unlikely. Hence, trend can be taken to be a proxy for some omitted variable or else one can go through the labors entailed in calculating a "real" interest rate and linking it to the monetary rate. Equation (4) is the familiar linear version of the *LM* curve.

The model was fitted to seasonally adjusted, deflated, quarterly data for the sample period 1953-65. Equations (3) and (4) were estimated by three-stage least squares whereas equation (2) was estimated by two-stage least squares in such a way that the product of the coefficients of  $L_t$  and  $C_{t-1}$  be equal to the coefficient of  $L_{t-1}$ .<sup>2</sup> It is well

known that the estimates of  $a$ ,  $b$ , and  $g$  are apt to be biased upward by single equation estimation techniques if the residuals have positive autocorrelation, although the results are not in yet for simultaneous equation estimation methods. At any rate, such bias as might exist is not terribly important for the argument advanced here. The results for the structural equations were:

$$(6) \quad C_t = -1.7 + 0.157Y_t + 0.047L_t \\ (0.8) \quad (0.013) \quad (0.031) \\ -0.035L_{t-1} + 0.750C_{t-1} \\ (0.022) \quad (0.057)$$

$$\bar{R}^2 = 0.95$$

$$(7) \quad I_t = 5.1 + 0.572(Y_t - Y_{t-1}) \\ (1.7) \quad (0.082) \\ -0.590r_t + 0.0365t + 0.787I_{t-1} \\ (0.324) \quad (0.018) \quad (0.056)$$

$$\bar{R}^2 = 0.92$$

$$(8) \quad M_t = 23.8 + 0.074Y_t - 1.989r_t \\ (9.3) \quad (0.017) \quad (0.564) \\ + 0.827M_{t-1} \\ (0.061)$$

$$\bar{R}^2 = 0.94$$

where the standard errors of the regression coefficients are shown directly below the coefficients themselves. Judging from the lag coefficients, the speed of adjustment is fairly slow and approximately equal in all three sectors. The signs of the coefficients are all as expected on a priori grounds, except perhaps for the trend variable, whose coefficient will be posited to be zero in what follows with little loss of generality.

## II. The Impact of Monetary Policy

In describing the movement from one equilibrium position to another after a change in the money supply, Tanner began from a position of initial equilibrium and

is the same used in estimating equation (7) and (8), which were estimated via the 3SLS method due to its generally more desirable statistical properties.

<sup>1</sup> The negative binomial lag distribution accounts for both lagged expectational and output adjustments, whereas the Koyck lag distribution considers only a single lagged adjustment.

<sup>2</sup> The necessity for the constraint is obvious from a casual glance at equation (2). The parameters were estimated by selecting different values for the lag coefficient  $a$  and estimating the remaining coefficients by the 2SLS technique in such a way that the constraint was satisfied. That lag coefficient which minimized the sum of squared residuals was the one selected. The set of predetermined variables used in estimating equation (6)

TABLE 1—TIME PATH FOR MONEY SUPPLY AND INTEREST RATE WITHOUT AND WITH ACCELERATOR AND LIQUID ASSETS

Period	Without Accelerator and Liquid Assets		With Accelerator and Liquid Assets	
	$\frac{M_t - M_0}{M_E - M_0}$	$\frac{r_t - r_0}{r_E - r_0}$	$\frac{M_t - M_0}{M_E - M_0}$	$\frac{r_t - r_0}{r_E - r_0}$
0	0.00	0.00	0.00	0.00
1	1.07	61.43	0.18	2.14
2	1.03	0.00	0.30	0.29
3	1.01	0.57	0.41	0.43
4	1.00	0.96	0.50	0.86
5	1.00	1.00	0.58	1.00
6	1.00	1.00	0.64	1.00

posited that some given change in aggregate demand was desired in the next period and all subsequent periods until a new equilibrium money supply and interest rate was attained. In doing so, he utilized the statistic

$$\frac{M_t - M_0}{M_E - M_0} \quad t = 1, 2, \dots, E$$

where  $M_0$  is the money supply at the initial equilibrium;  $M_E$  denotes the money supply at the final equilibrium; and  $M_t$  is the money supply required at the beginning of period  $t$  in order to obtain the desired value of the target variable during that period. The ratio would normally be expected to converge toward unity as the system approaches the new equilibrium. In addition, Tanner calculated the ratio

$$\frac{r_t - r_0}{r_E - r_0} \quad t = 1, 2, \dots, E$$

in order to estimate the reaction of the interest rate to changes in monetary policy. The pair of ratios were estimated for the model under two sets of conditions. In the first case, Tanner's model was approximated by specifying  $\beta_2$  and  $\lambda_1$  to be zero, hence eliminating the effects of the accelerator and the liquid assets variable in the consumption function from the two ratios. In the second case, the ratios were estimated with the ac-

celeration and liquid assets coefficients left in the model.

The results are presented in Table 1. If the accelerator and wealth effect in the consumption function are inoperative, the initial increase in the money supply must be sufficiently large to drive the interest rate way down in order to attain immediately the desired increase in aggregate demand. The lag structure then begins to take effect in the second period, and the required money supply gets smaller until it reaches its new equilibrium level in the fourth period. Meanwhile, the interest rate rises to its former level in the second period and then falls to its new lower equilibrium level. The important thing to note is the extreme flexibility required of the interest rate during the first two periods of the policy's operation.

If, on the other hand, the accelerator and liquid assets are included, an increase in outside money initially shifts the consumption function upward. Second, and more important, the increase in aggregate demand stimulated by the policy induced fall in the interest rate is magnified by a further accelerator induced increase in investment. The effect of both these results is to shift the  $IS$  curve temporarily to the right, increasing both aggregate demand and the interest rate. Because of these results and the subsequent larger impact multiplier and hence smaller required initial increase in the money supply, the necessary decrease in the interest rate is much smaller than in the more naive model. Inasmuch as the immediate impact of the accelerator peters out in the second period when aggregate demand has reached its new plateau, further increases in the money supply are necessary until a complete static equilibrium is reached.

### III. Conclusions

This brief communication has utilized a statistic suggested by Tanner in order to estimate the required time path of the money supply and the interest rate which must accompany a desired increase in aggregate demand. In the case of the simple  $IS-LM$  model Tanner's empirical conclusion that

monetary policy can be effective in the short run in spite of a long distributed lag in the product market was verified, although the necessary short-run change in the interest rate might be greater than could reasonably be expected. However, once the model is modified to include direct monetary effects in the product market and an investment accelerator, the desired expansion can be realized immediately with relatively small strain on the interest rate because of the shift in the *IS* curve and the larger impact multiplier for the money supply.

## REFERENCES

- J. E. Tanner, "Lags in the Effects of Monetary Policy: A Statistical Investigation," *Amer. Econ. Rev.*, Dec. 1969, 59, 794-805.
- D. P. Tucker, "Dynamic Income Adjustment to Money Supply Changes," *Amer. Econ. Rev.*, June 1966, 56, 433-49.
- A. Zellner, D. S. Huang, and L. S. Chau, "Further Analysis of the Short-Run Consumption Function with Emphasis on the Role of Liquid Assets," *Econometrica*, July 1965, 33, 571-81.

# Lags in the Effects of Monetary Policy: Reply and Some Further Thoughts

By J. ERNEST TANNER\*

There is now a rapidly accumulating body of work which suggests the lag in effect of monetary policy is short. The latest in this series is Paul Smith's model which implies that the impact effects of changes in the money stock are substantially greater than the long-run effects. If the effects of monetary policy on aggregate demand are not highly variable, such results would indicate that monetary policy is consistent with short-run stabilization requirements. It is the purpose of this note to comment briefly on the Smith paper and to show that the short lag found in recent papers is consistent with Milton Friedman's evidence, in his context, of both a long and variable lag.

## I. Evidence on the Short Lag

Up until 1966 when Donald Tucker published the theoretical argument for a short lag, it was widely held that the lag in effect of monetary policy was so great that active contra-cyclical monetary policy might aggravate rather than ameliorate fluctuations in economic activity. In strong support of Tucker's short lag hypothesis, in 1969 I wrote that my empirical results suggest "the impact effect of a change in the money supply is about equal to the long-run effect . . . (p. 801) (and) . . . the bulk of the effects on aggregate demand of monetary policy can occur within three to six months . . ." (p. 803). However, these conclusions were tempered with the observation that in order to get such quick short-run responses in aggregate demand, the model requires vigorous overshooting in the rate of interest. In his comment, Smith incorporates liquid assets into the consumption function and an accelerator into the investment relationship of the *IS-LM* model and finds that the short-run effects are even larger than

those I found and can occur without the undesirable large swings in the rate of interest. These extensions are important contributions and his results indicate that monetary policy would be ideal for short-run stabilization requirements.

Unfortunately, in making these extensions, Smith imposed the geometric lag distribution upon each structural equation. Because this distribution imposes the constraint that the lag coefficients decline geometrically, the empirical model does not allow the effects of monetary policy on income to build for the first few quarters and then taper away. In contrast to the more intuitive appealing result I obtained on the impact effect being about equal in magnitude to the long-run effect but substantially less than the intermediate six to nine months effect, Smith's constraint does not allow income to continue to build in these intermediate periods before adjusting back. Rather, the Smith constraint requires that the dynamic path of income approach equilibrium asymptotically following the first period response. Consequently, in the dynamic model which excludes the accelerator and liquid assets variables, Smith finds that the impact effect is over 90 percent of the long-run effect with equilibrium being reached by the fourth period. The alternative formulation which includes the accelerator and liquid assets variable suggests the impact effect of a change in the money supply is significantly greater than the long-run effect but the effects begin tapering away by the second quarter. In either case, following the initial shock, income approaches equilibrium directly.

Had the more general negative binomial lag distribution been used, the model could have been flexible enough to permit the effects on income to continue building for a few quarters before tapering away. In light of other empirical evidence obtained from a

\* Associate professor of economics, Tulane University.

TABLE 1—DYNAMIC PROPERTIES OF INCOME ADJUSTMENT FOLLOWING A  
ONCE AND FOR ALL CHANGE IN THE MONEY SUPPLY IN PERIOD ONE AS  
GIVEN BY THE STATISTIC  $(Y_t - Y_0)/(Y_E - Y_0)$

Time Period (Quarters)	Direct Model	Negative Binomial <i>IS-LM</i> Model	Geometric <i>IS-LM</i> Model 1	Geometric <i>IS-LM</i> Model 2
$t \leq 0$	0.0	0.0	0.0	0.0
1	.50	1.05	.94	5.55
2	1.23	1.11	.97	3.33
3	1.82	1.19	.99	2.44
4	1.95	1.15	1.00	2.00
5	1.70	1.12	1.00	1.72
6	1.30	1.07	1.00	1.56
7	1.00	1.04	1.00	1.41
.	.	.	.	.
.	.	.	.	.
Equilibrium	1.00	1.00	1.00	1.00

*Note:* In constructing this table, it is assumed that the money supply is constant and equal to  $M_0$  for a period long enough so that income is initially at equilibrium. At time  $t=1$ , the money supply increases to  $M_E$  and remains there for all future time. By using the statistic  $(Y_t - Y_0)/(Y_E - Y_0)$  where  $Y_0$  is the initial equilibrium level of income corresponding to the initial money stock  $M_0$ ,  $Y_E$  is the final equilibrium level of income corresponding to the new money stock  $M_E$ , and  $Y_t$  is the implied income level in any period  $t$  resulting from a change in the money stock from  $M_0$  to  $M_E$  in period 1, the table shows the dynamic adjustment path of income as implied by four recent econometric models. The Direct Model column was calculated from Michael W. Keran, Chart III, p. 13 for the 1919–1969 period. The Negative Binomial *IS-LM* Model column was calculated from my original article. The Geometric *IS-LM* Model 1 column was calculated from Smith's results, without accelerator and liquid assets, whereas the Geometric *IS-LM* Model 2 column was calculated from Smith's results, with accelerator and liquid assets.

direct regression of changes in income on current and past changes in the money stock, such flexibility would have been desired. Table 1 compares the results of four different estimates of the dynamic path of income to a new equilibrium resulting from a once and for all change in the money stock. The first column shows a typical result of the direct least squares approach indicating that the monetary effects on income continue to build for three or four quarters and then become negative. In column 2, a similar dynamic behavior is implied by my model published in 1969<sup>1</sup> and the implications of

Smith's model are given in columns 3 and 4. As the table shows, the Smith model indicates that income adjusts directly towards equilibrium after the initial shock whereas the other two models indicate that the effect on income builds for three or four quarters before adjusting towards equilibrium. But, because the purpose of Smith's comment is

<sup>1</sup> I would like to thank Herschel Grossman for pointing out that the technique of augmented least squares which will theoretically give consistent parameter estimates of distributed lagged relationship even in the presence of serial correlation, was incorrectly applied in that paper. The serial correlation in the commodity demand equation appears to be of the form  $W_t = V_t - \beta V_{t-1}$  where  $W$  is a truly random disturbance,  $V$  is the equation error

term and  $\beta$  is a structural lag parameter. To rid the function of this systematic influence, consistent estimates of  $V_t$  and  $V_{t-1}$  were incorrectly introduced into the estimation equation instead of correctly introducing  $V_{t-1}$  alone. Such a procedure produces inconsistent parameter estimates because  $W_t$  is correlated with  $V_t$ . However, because the lag in effects within the *IS-LM* model used depends roughly upon whether the commodity or monetary sector responds more quickly and because the problem of autocorrelation in the monetary sector was avoided by assuming money to be exogenous, it is likely that the biases in the estimated lag coefficients are in the same direction for both sectors and therefore, largely offsetting.

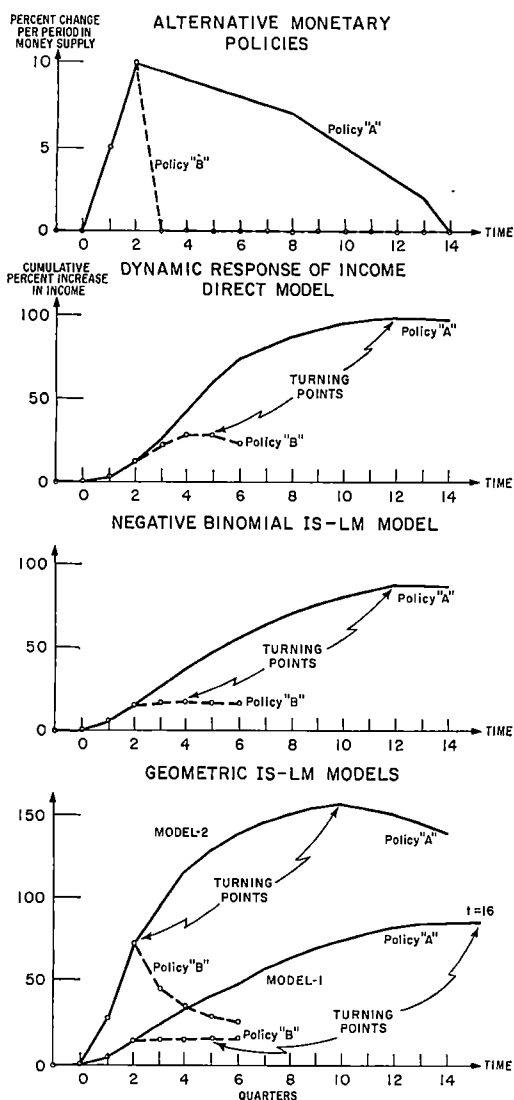


FIGURE 1

to show that the large fluctuations required of the interest rate in the first two periods in my formulation are not required when liquid assets and an accelerator term are allowed to operate, it is likely that the geometric lag distributions are not too bad approximations. However, if the model is used to say something about the adjustment of income and the rate of interest in later periods, more general lag distributions would be preferred.

## II. The Variable Lag Hypothesis

All the models summarized in Table 1 indicate a short lag between changes in the money supply and corresponding changes in aggregate demand. Therefore monetary policy has seemingly been vindicated. However, Friedman has argued that the lag is highly variable and, therefore, contracyclical monetary policy is not vindicated because "monetary actions in large measure introduce a random disturbing element into economic affairs" (1961, p. 448). By comparing the timing of peaks and troughs in the rate of change in the money supply with peaks and troughs in general business activity as dated by the National Bureau Business cycle reference points<sup>2</sup> he finds that money leads income by an average of 18 months at peaks and an average of 12 months at troughs (see Friedman and Anna Schwartz, p. 38). More importantly, Friedman finds that these leads are highly variable with a standard deviation of 6 to 7 months (about two quarters).

By using Friedman's method, we also obtain long and variable lags even though the dynamic relationship between money and income is assumed to be exact and the lag is assumed to be short. Consider, for example, the effect on the turning points of income resulting from two hypothetical monetary policies as plotted in the top panel of Figure 1.

In this figure it is assumed that income is at equilibrium corresponding to the money supply at time  $t=0$ . In period  $t=1$ , the money supply increases by 5.0 percent and in period  $t=2$ , the money supply increases an additional 10 percent. Beginning in period  $t=3$ , monetary policies A and B begin to differ. In policy B the money stock remains constant from period  $t=3$  onwards whereas in policy A the money supply con-

<sup>2</sup> Such a method is very inefficient in that it ignores most of the information contained in the money supply series. Consequently, Friedman's method leads to an extremely large estimated variance in the lag. A proper specification of the Friedman approach would take account of, in addition to the turning point, the changes in the money supply both before and after the turning point as well as other factors. A similar point has also been made by Thomas Mayer, pp. 336-37.

tinues to increase but at slower and slower rates. Using the dynamic relationships between money and income given in Table 1, the time paths of income for the two policies are plotted in the lower three panels of Figure 1 for each of the models. In this figure, the cumulative percentage increase in income from  $t=0$  is plotted on the vertical axis and time is plotted on the horizontal axis.<sup>3</sup> The figure shows that, in each model, the lack of growth in the money supply in policy B causes income to turn down considerably before income turns down with policy A which allows money to continue growing but at reduced rates. The direct model gives a variable lag of seven quarters, the negative binomial *IS-LM* model gives a variable lag of eight quarters whereas Smith's models given variable lags of eight and eleven quarters.

As the panels in Figure 1 graphically illustrate, looking only at turning points in the money supply could be very misleading. Although these examples do not prove that the lag is not highly variable, a world with substantially no lags in the effects of monetary policy and with highly predictable effects of money supply changes on income would

look as if it had a long and variable lag if we looked solely at turning points. The point is that the time path of income depends more upon the magnitude and duration of an expansion (or contraction) in the money supply than its turning point. Turning points alone tell us very little about the length or variability of the lags in monetary policy.

#### REFERENCES

- M. Friedman, "The Lag in the Effect of Monetary Policy," *J. Polit. Econ.*, Oct. 1961, 69, 447-66.
- and A. Schwartz, "Money and Business Cycles," *Rev. Econ. Statist.*, Feb. 1963 supp., 45, 32-64.
- M. W. Keran, "Monetary and Fiscal Influences on Economic Activity—The Historical Evidence," *Review*, Federal Reserve Bank of St. Louis, Nov. 1969, 5-24.
- T. Mayer, "The Lags in the Effects of Monetary Policy: Some Criticisms," *Western Econ. J.*, Sept. 1967, 5, 324-42.
- P. E. Smith, "Lags in the Effects of Monetary Policy: Comment," *Amer. Econ. Rev.*, Mar. 1972, 62, 230-33.
- J. E. Tanner, "Lags in the Effects of Monetary Policy: A Statistical Investigation," *Amer. Econ. Rev.*, Dec. 1969, 59, 794-805.
- D. P. Tucker, "Dynamic Income Adjustment to Money Supply Changes," *Amer. Econ. Rev.*, June 1966, 56, 433-49.

<sup>3</sup> The cumulative percentage increases in income figures are derived on the assumption that the long run velocity of money is constant, i.e., the new equilibrium velocity after the experiment will be the same as velocity before the experiment.

## ERRATA

### Optimal Taxation and Public Production: I

By PETER A. DIAMOND AND  
JAMES A. MIRRLEES

In our article, there is an error on page 18 of the March 1971 issue of the *Review*, brought to our attention by John A. Nordin. The social welfare function should be

$$\frac{-1}{x_1^2 y_1} - \frac{1}{x_2 y_2^2}$$

### Toward a Theory of Nonprofit Institutions: An Economic Model of a Hospital

By JOSEPH P. NEWHOUSE

I wish to correct an error in my article published in the March 1970 issue of the *Review*.

The last paragraph on page 68 states that a constrained quantity-quality maxi-

mizing firm will carry production to the point where marginal revenue product equals marginal factor cost. This is incorrect. The correct marginal condition is that the marginal rate of technical substitution between any two factors will equal their price ratios. This error was pointed out by David Kamerschen, John Rafferty, Richard Wallace, and John Hoag. The paragraph can be corrected as follows:

We are interested in using this model to assess the effect of the hospital's nonprofit status upon efficiency. First, note that this model implies least-cost production insofar as the decision maker pursues his maximization goals. Suppose that the marginal rate of substitution between two factors exceeds their price ratio. A profit-maximizing firm would substitute one factor for another to the point of equality, thereby achieving a socially optimal allocation of factors. A constrained quantity-quality maximizer will also substitute to the point of equality. He can use the "profit" the profit-maximizing firm would have gained to shift his quantity-quality trade off curve out and thereby reach a higher indifference curve. When the equalities hold for all factors, the tradeoff curve cannot be shifted out any more.

# ANNOUNCEMENT

## NOTICE TO ALL GRADUATE DEPARTMENTS

The December 1972 issue of the *Review* will carry the sixty-ninth list of doctoral dissertations in political economy in American universities and colleges. The list will specify doctoral degrees conferred during the academic year terminating June 1972. This announcement is an invitation to send us information for the preparation of the list. This announcement supercedes and replaces a letter which was sent annually from the managing editor's office.

The *Review* will publish in its December 1972 issue the names of those who will have been awarded the doctoral degree since June 1971, the titles of their dissertations, and, if possible, a brief (75-word) summary of the dissertation.

By June 30, please send us this information on 3×5 cards, conforming to the style shown below, one card for each individual. Please indicate by a classification number in the right-hand corner the field in which the thesis should be classified. The classification system is that used by the *Journal of Economic Literature* and printed in every issue.

Name: <u>LAST NAME IN CAPS: First Name, Initial</u>	<i>JEL</i> Classification No. _____
Institution Granting Degree: _____	
Degree Conferred (Ph.D. or D.B.A.) _____ Year _____	
Dissertation Title: _____	
Summary	
(75-word maximum, or first 75 words will be printed)	
Summary may be completed on back of this card or on new card which should be stapled to this.	

When degrees in economics are awarded under different names, such as Business Administration, Public Administration, or Industrial Relations, candidates in these fields whose training has been *primarily in economics* should be included.

# NOTES

## *Notice to Members and Subscribers*

Members of the American Economic Association receive preregistration materials for the annual meetings together with the Association's publications. Normally subscribers do not receive preregistration materials because most subscriptions are held by institutions. In some cases, however, individual subscribers may wish to receive preregistration materials. These may be obtained by writing to the Secretary's office, Nashville, Tennessee.

A Summer Institute for college teachers of economics will be held at Brown University from June 18 to July 28, 1972. The institute is supported by a grant from the National Science Foundation. The program is designed to acquaint teachers in undergraduate institutions with recent developments and policy applications of micro-economic theory. For application forms and further information, write Professor Mark B. Schupack, NSF Institute in Economics, Department of Economics, Brown University, Providence, Rhode Island 02912.

A Summer Institute in Urban Economics, for college and university teachers of economics, will be held at Stanford University, July 30-August 26, 1972. The institute is sponsored jointly by Stanford University and by the Inter-University Committee on Urban Economics, and is supported by the National Science Foundation.

Twenty-five participants will be selected. The Institute will expose the participants to the subject matter and methodology of urban economics and will prepare participants for teaching in this field. For application forms and further information, write Professor Benton F. Massell, NSF Institute in Urban Economics, Food Research Institute Building, Stanford University, Stanford, California 94305.

The National Tax Association announces the 1971 award winners in the first annual competition for outstanding doctoral dissertations in government finance and taxation. The award was won by Michael Jay Boskin of University of California, Berkeley, with his entry, "The Effects of Taxes on the Supply of Labor with Special Reference to Income Maintenance Programs." Honorable mention awards were won by Robert P. Inman of Harvard University, "Four Essays on Fiscal Federalism," and Nancy D. Sidhu of the University of Illinois, "The Effects of Changes in Sales Tax Rates on Retail Prices."

Information on the 1972 award competition may be obtained from Professor James A. Papke, department of economics, Krannert Graduate School of Industrial Administration, Purdue University, Lafayette, Indiana 47906. The 1972 selection committee consists of Pro-

fessors George Break, Cary Brown, Arthur Lynn, James Papke, and Clyde Reeves. The award is \$1,500 and publication of a summary of the dissertation. The Honorable Mention awards are \$500 and publication.

The Committee on International Exchange of Persons announces that applications for senior Fulbright-Hays awards for lecturing and research during 1973-74 in about 80 foreign countries will be accepted in the spring of 1972. Scholars who are U.S. citizens and have a doctorate or college teaching experience are invited to indicate their interest by completing a simple registration form, available on request from *Senior Fulbright-Hays Program*, 2101 Constitution Ave., Washington, D.C. 20418. Registrants will receive the detailed announcement of available awards as soon as it is issued in the spring, in time to consider the possibilities and to apply before the closing date. July 1, 1972 is the deadline for applying for research awards and it is the suggested date for filing for lectureships.

Senior Fulbright-Hays awards generally consist of a maintenance allowance in local currency to cover normal living costs of the grantee and family while in residence abroad, and round-trip travel for the grantee (transportation is not provided for dependents). For lecturers going to most non-European countries, the award includes a dollar supplement, subject to the availability of funds, or carries a stipend in dollars and foreign currency, the amount depending on the assignment, the lecturer's qualifications, salary, and other factors.

The second winner of the Columbia University Prize in American Economic History is William G. Whitney, assistant professor of economics at the University of Pennsylvania, for his manuscript "The Structure of the American Economy in the Late Nineteenth Century." The award of \$1000 and publication by the Columbia University Press is intended to encourage clarity and persuasiveness in economic analysis. Eligible for it are book-length manuscripts dealing with any aspect of American economic history broadly conceived. Manuscripts for consideration for the next award should be sent before June 30, 1972, to History Editor, Columbia University Press, 562 West 113th Street, New York, New York 10025.

A special fellowship in population and economics has been established in the name of Ritchie H. Reed. The fellowship was announced jointly by The Population Council, which will administer it, and the Commission on Population Growth and the American Future, where Reed was employed as a government economist. The

Population Council is located at 245 Park Avenue, New York, New York 10017.

A limited number of Public Health Service traineeships are available for a doctoral training program in medical care organization and administration. Stipends range from \$3,000 to \$4,200 depending on previous experience, plus a dependency allowance and tuition. In addition, a number of teaching assistantships are available. In return for 15 hours of work a week, they provide full tuition and fees plus a stipend of \$2,800 per year. Students may choose to develop a special expertise in such areas as operations research, finance, health economics, organizational theory, medical sociology or political science, along with developing a basic familiarity with the problems of medical care organization. For further information and application forms, write to Program Coordinator, Sloan Institute of Hospital and Health Services Administration, Malott Hall, Cornell University, Ithaca, New York 14850.

The Mathematical Social Science Board and the Committee on the Undergraduate Program in Mathematics are seeking interesting problems of illustrative examples from each of the social sciences whose solutions and study make use of ideas and techniques from one or more of the following topics in undergraduate mathematics: sets and relations; differential and integral calculus; matrices and linear algebra; and probability.

We propose to collect such examples into a book to be used by mathematics teachers and students as a source of (1) current social science applications of mathematics and of (2) material for textbook and classroom exercises to illustrate how topics in collegiate mathematics arise in a social science context. We also plan to include annotated bibliographies of articles and books involving applications of mathematics to the various social sciences. Please send contributions to: CUPM-MSSB Project, P.O. Box 1024, Berkeley, California 94701.

The 12th annual Washington Conference on Business-Government Relations sponsored by The American University Center for the Study of Private Enterprise will be held on April 3, 4, 1972 at the Shoreham Hotel, Washington, D.C. The theme will be "A Dialogue with the EPA," and the subtheme, "A Discussion of the Present and Future Role of the Environmental Protection Agency in a Dynamic Business-Government Relationship." For further information, contact Dr. Jimmy D. Johnson, Director, The American University, Center for the Study of Private Enterprise, Massachusetts and Nebraska Avenues, Washington, D.C. 20016.

### *Deaths*

Lewis W. Adams, professor of economics, School of Commerce, Economics, and Politics, Washington and Lee University, Apr. 3, 1971.

Roy E. Cameron, professor emeritus, San Diego State College, October 7, 1971.

Elmer Fagan, Stanford University, June 16, 1971.

Eliot Jones, Stanford University, Oct. 17, 1971.

Ritchie H. Reed, director of economic research, Commission on Population Growth and the American Future, Washington, Oct. 1, 1971.

Leon D. Smith, assistant professor of business administration, Fisk University, May 20, 1971.

George W. Thatcher, professor of economics, Miami University, May 9, 1971.

James M. Waller, retired visiting professor of economics, University of Virginia, Aug. 9, 1971.

### *Retirements*

Corwin D. Edwards, professor of economics, University of Oregon, June 1971.

William H. McPherson, professor emeritus, University of Illinois, Urbana, Aug. 31, 1971.

George L. Mehren, professor of agricultural economics; economist, Experiment Station, Giannini Foundation, Jan. 1971.

Alden J. Plumley, professor of economics, University of Nevada, June 1972.

Randall W. Tucker, associate professor of economics, Trinity College, Feb. 1, 1972.

### *Visiting Foreign Scholars*

John Bates, University of Nottingham: visiting professor of economics, University of California, San Diego, winter and spring quarters, 1972.

Ernst W. Boehm, University of Melbourne: visiting scholar, department of economics, University of Pittsburgh, Jan. 1, 1972.

Marwan Ghandour, Lebanon: visiting scholar, department of economics, Stanford University.

Bernard Gotschalk, Germany: visiting scholar, department of economics, Stanford University.

Raymond F. D. Hutchings, Royal Institute of International Affairs, London: visiting professor of economics, University of Maryland, Feb.-June 1972.

Ruud J. H. Jansen: research assistant, University of Alabama.

Shigeki Kanada, Tohoku University, Japan: visiting scholar, department of economics, University of Colorado, Jan. 1-Feb. 28, 1972.

Sadao Katayama, Japan: visiting scholar, department of economics, Stanford University.

T. O. M. Kronsjö, University of Birmingham: visiting professor of economics, Purdue University, Feb.-June 1972.

Tzong Biau Lin, Hong Kong: visiting scholar, department of economics, Stanford University.

Matthew T. G. Meulenbergh, Agricultural University at Wageningen, The Netherlands: visiting professor, School of Agricultural Economics and Extension Education, University of Guelph, May-Aug. 1972.

E. J. Mishan, London School of Economics: visiting professor of economics, University of Virginia, second part of spring semester, 1972.

Miroslav Petrović, Belgrade University, Yugoslavia: visiting professor, department of economics, Western Michigan University, Jan.-July 1972.

C. Rangarajan, Indian Institute of Management: visiting professor of finance, Graduate School of Business Administration, New York University.

Joan Robinson, Cambridge University: visiting professor of economics, McGill University, first term, 1971-72.

Yoshio Shimizu, Konan University, Japan: visiting scholar, department of economics, University of Colorado, Aug. 1970-Mar. 31, 1972.

Anthony P. Thirlwall, University of Kent at Canterbury, England: visiting research economist, Princeton University, 1971-72.

### *Promotions*

Dale Adams: professor of agricultural economics, Ohio State University.

Donald E. Baer: assistant professor of economics, University of Illinois at Chicago Circle.

Marion S. Beaumont: associate professor of economics, California State College at Long Beach.

Donald M. Bellante: assistant professor of economics, Auburn University.

Charles A. Berry: associate professor of economics, University of Cincinnati.

Robert D. Britt: associate professor of economics, West Virginia University.

Liberato M. Cacace: economist, financial statistics division, Federal Reserve Bank of New York.

Burnham O. Campbell: professor of economics, University of Hawaii.

Laurits Christensen: associate professor of economics, University of Wisconsin, Madison.

James W. Christian: professor of economics, Iowa State University.

Elchanan Cohn: associate professor of economics, Pennsylvania State University, July 1, 1971.

J. Malcolm Dowling, Jr.: associate professor, department of economics, University of Colorado, fall 1971.

George R. Dreese: associate professor of economics, West Virginia University.

Kenneth G. Elzinga: associate professor, department of economics, University of Virginia, Sept. 1, 1971.

Thomas P. Enger: associate professor of economics, St. Olaf College, Sept. 1, 1971.

Bernard Erven: associate professor of agricultural economics, Ohio State University.

Marianne A. Ferber: assistant professor of economics, University of Illinois, Urbana, Sept. 1, 1971.

Louis Fier: professor, department of economics, Brooklyn College of the City University of New York, Jan. 1, 1972.

Frances V. Flanagan: assistant professor of economics, University of Illinois at Chicago Circle.

Thomas G. Fox: associate professor of economics, Pennsylvania State University, July 1, 1971.

Murray A. Goldberg: instructor of finance, University of Illinois at Chicago Circle.

Christopher Green: associate professor of economics, McGill University, Sept. 1, 1971.

David Hahn: associate professor of agricultural economics, Ohio State University.

Jagdish Handa: associate professor of economics, McGill University, Sept. 1, 1971.

James S. Hanson: assistant professor of economics, Wesleyan University, July 1, 1971.

Jerry G. Hunt: associate professor of business administration, New Mexico State University.

Charles Ingraham: professor of agricultural economics, Ohio State University.

Sol Jacobson: professor, department of economics, Brooklyn College of the City University of New York, Jan. 1, 1972.

Allen C. Kelley: professor of economics, University of Wisconsin, Madison.

K. C. Kogiku: professor, department of economics, University of California, Riverside, July 1971.

J. David Lages: professor of economics, Southwest Missouri State College, Sept. 1, 1971.

J. K. Lee: associate professor of economics, University of Cincinnati.

Peter Lindert: associate professor of economics, University of Wisconsin, Madison.

Charles R. Link: assistant professor, department of economics, University of Delaware, Feb. 1971.

Edward O. Lutz: professor, department of economics, Brooklyn College of the City University of New York, Jan. 1, 1972.

Michael S. McPherson: instructor of economics, University of Illinois at Chicago Circle.

Patrick C. Mann: associate professor of economics, West Virginia University.

J. Barry Mason: professor, department of marketing, University of Alabama, June 1971.

Elroy Mestre: professor of economics, State University College of New York at Oneonta.

Walter Miklius: professor of economics, University of Hawaii.

John C. Narver: professor, School of Business Administration, University of Washington, July 1, 1971.

Seiji Naya: professor of economics, University of Hawaii.

William H. Newell: assistant professor in economics, St. Olaf College, Sept. 1, 1971.

James R. Prescott: professor of economics, Iowa State University.

Jack L. Robinson: professor of economics, University of Oklahoma.

Hyman Sardy: professor, department of economics, Brooklyn College of the City University of New York, Jan. 1, 1972.

Stephen J. Schmutte: assistant professor, department of economics, Wabash College, Sept. 1971.

Nicholas W. Schrock: associate professor, department of economics, University of Colorado, fall 1971.

Joseph J. Seneca: associate professor, department of economics, Rutgers—The State University, July 1971.

Roger P. Sherman: professor, department of economics, University of Virginia, Sept. 1, 1971.

Fred S. Silander: professor of economics, DePauw University, July 1971.

Lois Simonds: associate professor of agricultural economics, Ohio State University.

Kenneth R. Smith: associate professor of economics, University of Wisconsin, Madison.

Allan D. Spritzer: assistant professor of management, University of Alabama, Aug. 1971.

Dennis R. Starleaf: professor of economics, Iowa State University.

J. Kirker Stephens: associate professor of economics, University of Oklahoma.

Vern Vandemark: associate professor of agricultural economics, Ohio State University.

Charles E. Vinson: associate professor, finance department, College of Commerce and Business Administration, University of Alabama, Sept. 1, 1971.

Thomas J. Weiss: associate professor, department of economics, University of Kansas.

Marina vN. Whitman: professor, department of economics, University of Pittsburgh.

Ray D. Whitman: assistant professor of economics, University of Maryland, fall 1971.

Joseph R. Zandarski: professor, department of business and economics, University of Scranton, fall 1971.

### *Administrative Appointments*

Moses Abramovitz: chairman, department of economics, Stanford University.

Leon Applebaum: professor; chairman, social science division, University of Wisconsin, Parkside.

Andrew Jong-Chong Au: chairman, department of economics, Millersville State College, Aug. 1971.

Morton S. Baratz, Boston University: vice chancellor of academic affairs, University of Maryland, Baltimore County, Jan. 1972.

T. Bruce Birkenhead: professor; dean, School of Social Sciences, Brooklyn College of the City University of New York, Jan. 1, 1972.

Ted R. Brannen, University of Houston: dean, University of Southern California School of Business Administration, Jan. 1, 1972.

Robert Campbell: chairman, department of economics, University of Oregon, Sept. 1971.

Alan Carlin, The Rand Corporation: director, Implementation Research Division, Office of Research and Monitoring, Environmental Protection Agency, Washington, Aug. 1971.

John F. Due: chairman, department of economics, University of Illinois, Urbana, Sept. 1, 1971.

Joseph R. Guerin: chairman, department of economics, St. Joseph's College, Sept. 1971.

Mark R. Killingsworth: acting chairman, department of economics and business administration, Fisk University, Sept. 1971.

David Kresge: director of doctoral programs, Graduate School of Business Administration, New York University.

Charles N. Lanier: acting chairman, department of economics, University of Delaware, 1971-72.

J. Barry Mason: director of research and special programs, College of Commerce and Business Administration, University of Alabama.

Hans Mueller: professor; chairman, department of economics, Middle Tennessee State University.

Benjamin M. Perles: dean, College of Business, Eco-

nomics, and Government, University of Alaska, July 1, 1971.

J. Donald Phillips: chairman, department of management, University of Alabama, Aug. 1971.

John Rapp: professor; chairman, department of economics, MacMurray College.

James N. Rosse: associate professor; vice chairman, department of economics, Stanford University.

Gary W. Sorenson: chairman, department of economics, Oregon State University, Sept. 15, 1971.

M. Richard Sussman, University of Georgia: chairman, department of economics and business administration, State University of New York College at Fredonia, Aug. 1971.

Izumi Taniguchi: chairman, department of economics, Fresno State College, Sept. 1, 1971.

Benjamin J. Taylor: chairman, department of economics, University of Oklahoma.

John A. Tomaske: associate dean of academic planning, graduate studies, and research, California State College, Los Angeles, Sept. 15, 1971.

G. I. Trant: University of Guelph: director general, economics branch, Canada Department of Agriculture, Ottawa, Jan. 1, 1972.

Charles Waldauer: head, department of economics, PMC Colleges.

T. K. Warley: director, School of Agricultural Economics and Extension Education, University of Guelph, July 1, 1971.

Thomas A. Yancey: associate dean, College of Commerce and Business Administration, University of Illinois, Urbana, Sept. 1, 1971.

### *Appointments*

Irma Adelman, Northwestern University: senior economist, Development Research Center, World Bank, Sept. 1971.

Ifthikhar Ahmed: research associate, department of economics, Iowa State University.

Edward L. Allen: deputy assistant secretary, International Economic Research and Analysis, U.S. Department of Commerce.

Emily Andrews: economist, domestic research division, Federal Reserve Bank of New York.

R. Keith Aufhauser: assistant professor, department of economics, Queens College of the City University of New York.

Rolf Auster, Northwestern University: assistant professor of finance, University of Illinois at Chicago Circle.

Louis Baeriswyl: staff member, management sciences department, The Rand Corporation, June 1971.

Anthony Ballman, University of Missouri: production management department, Moorman Manufacturing Company, Quincy, Illinois, Aug. 1971.

Robert L. Batterham: assistant professor of agricultural finance, School of Agricultural Economics and Extension Education, University of Guelph, Sept. 1, 1971.

Richard A. D. Beck: instructor, department of economics, Iowa State University.

Joe A. Bell: instructor of economics, Southwest Missouri State College, Sept. 1, 1971.

Sidney Bennett, Georgia State University: lecturer, department of marketing, University of Alabama, fall 1971.

James Bicksler, Rutgers University: visiting associate professor of finance, University of Illinois at Chicago Circle.

Malcolm Blackie, University of Missouri: research fellow, University of Nottingham, England, Aug. 1971.

Daniel Blake: assistant professor, department of economics, San Fernando Valley State College, Sept. 1, 1971.

Bill Blakeslee, University of Missouri: vice president for fluid milk marketing, Mid America Dairymen, Inc., Springfield, Missouri, Dec. 1971.

Alan S. Blinder: assistant professor, department of economics, Princeton University, Sept. 1971.

Charles R. Blitzler, Stanford University: economist, Development Research Center, World Bank, Sept. 1971.

Alan E. Boese: assistant professor of economics, Virginia State College.

H. Woods Bowman, Federal Reserve Bank of Chicago: assistant professor of economics, University of Illinois at Chicago Circle.

Michael E. Bradley: assistant professor of economics, University of Nevada, Sept. 1, 1971.

Robert Brogan: assistant professor of economics, State University College of New York at Oneonta.

Phillip D. Brooks: lecturer, department of economics, University of Nevada, Sept. 1971.

Donald M. Brown: instructor, department of economics, Miami University, Middletown, fall 1971.

Sandra C. Christensen: assistant professor, department of economics, University of Maryland, 1971-72.

Peter B. Clark, Massachusetts Institute of Technology: senior economist, Development Research Center, World Bank, Sept. 1971.

Edward V. Daley: assistant professor, department of economics, Middle Tennessee State University.

Dennis N. DeTray, University of Chicago: staff member, economics department, The Rand Corporation, Oct. 1971.

Ahmed El-Safty, Massachusetts Institute of Technology: assistant professor, department of economics, Eastern Michigan University.

Dawn E. Elvis: assistant professor of economics, Fisk University.

Okon J. Essien: assistant professor of economics, Fisk University.

Keith D. Evans: associate professor, department of economics, San Fernando Valley State College, Sept. 1, 1971.

Eva R. Ewing: staff member, economics department, The Rand Corporation, Sept. 1971.

J. Michael Fitzmaurice: assistant professor, department of economics, University of Maryland, 1971-72.

Richard L. Floyd: instructor, department of economics, Iowa State University.

Gwen A. Fountain, University of Michigan: assistant professor, department of economics, Eastern Michigan University.

Milton Friedman, University of Chicago: visiting professor of economics, University of Hawaii, spring 1972.

Thomas F. Funk: assistant professor of agribusiness, School of Agricultural Economics and Extension Education, University of Guelph, Aug. 1971.

Arthur Gandolfi: economist, domestic research division, Federal Reserve Bank of New York.

William Gasser: economist, foreign research division, Federal Reserve Bank of New York.

Robert F. Glenn, University of Missouri: assistant professor, Southwest Missouri State College, Sept. 1971.

E. C. Gray: associate professor of resource economics, School of Agricultural Economics and Extension Education, University of Guelph, May 1, 1971.

Michael Grossman, University of Chicago: research associate, National Bureau of Economic Research, Sept. 1971.

Robert Hammer: senior economic advisor, data processing division, IBM Corporation.

John R. Hanson II: economist, foreign research division, Federal Reserve Bank of New York.

Donald K. Hargreaves: economist, balance of payments division, Federal Reserve Bank of New York.

James A. Hefner, Clark College: visiting research economist and lecturer in public affairs, Princeton University, 1971-72.

H. Robert Heller: professor of economics, University of Hawaii.

John S. Hodgson: assistant professor of economics, University of Oklahoma.

Robert C. Hsu, Michigan University: assistant professor of economics, Clark University, Sept. 1, 1971.

Glenn R. Hueckel, University of Wisconsin: assistant professor of economics, Purdue University, Feb. 1972.

John F. Hurley: associate professor of economics, Virginia State College.

Kevin Hurley: economist, domestic research division, Federal Reserve Bank of New York.

Louis C. Jacoby: assistant professor in economics, St. Olaf College, Sept. 1, 1971.

Louis J. James: assistant professor, finance department, College of Commerce and Business Administration, University of Alabama, Sept. 1, 1971.

Karen H. Johnson: instructor in economics, Wellesley College.

John F. Johnston: instructor, department of economics, University of Delaware, Sept. 1971.

Kiyoshi Kawahito: assistant professor, department of economics, Middle Tennessee State University.

Susan S. Kellar, University of Cincinnati: instructor, department of economics, Miami University, Middletown, fall 1971.

Eric Kierans: visiting professor of economics, McGill University, second term, 1971-72.

Han Kim, Stanford University: economist, Development Research Center, World Bank, Sept. 1971.

William C. Kleiner, Michigan State University: assistant professor, department of economics, Western Illinois University.

August W. Kanauber: assistant professor of economics, Manhattan College of the City University of New York, Sept. 1, 1971.

Roger Kubarych: economist, balance of payments division, Federal Reserve Bank of New York.

Allen W. Lacy: assistant professor of economics, Auburn University.

Helen F. Ladd: instructor in economics, Wellesley College.

William R. Latham: instructor, department of economics, University of Delaware, Sept. 1971.

William Lazarow: assistant professor of accounting, Brooklyn College of the City University of New York, Sept. 1, 1971.

Victor D. Lippit: assistant professor, department of economics, University of California, Riverside, July 1971.

Patrick R. Liverpool: instructor in business administration, Fisk University.

Constantino Lluch, Louvain University: economist, Development Research Center, World Bank, Sept. 1971.

Ashley Lovell, University of Missouri: assistant professor, Tarleton State College, Aug. 1971.

Harold Loyd, University of Missouri: assistant professor, Abraham Baldwin College of Agriculture, Aug. 1971.

Richard O. Lundquist: department of economics and business administration, State University of New York College at Fredonia, Aug. 1971.

Francis E. McCormick: assistant professor, department of economics, University of California, Riverside, July 1971.

John McDonald, Yale University: instructor of economics, University of Illinois at Chicago Circle.

Charles R. McKnew, Jr.: assistant professor of economics, West Virginia University.

Dilip B. Madan: assistant professor, department of economics, University of Maryland, 1971-72.

Robert Maisel: instructor of economics, University College, New York University, Sept. 1971.

Michael J. Maran: instructor of economics, Brooklyn College of the City University of New York, Sept. 1, 1971.

James Marchand: assistant professor, department of economics, San Fernando State College, Sept. 1, 1971.

Larry J. Martin: assistant professor of marketing, School of Agricultural Economics and Extension Education, University of Guelph, Jan. 1, 1972.

Jacob Metzger: assistant professor of economics, Northeastern Illinois University, fall 1971.

Arnold B. Moore: senior staff, economics department, The Rand Corporation, June 1971.

Noel K. Morris, University of Missouri: area farm management agent, University of Missouri Extension Division, Aug. 1971.

Rodney J. Morrison: associate professor of economics, Wellesley College.

Irene M. Moszer: associate professor of economics, Virginia State College.

Paul G. Munyon: instructor in economics, Wellesley College.

Steve Newcom, University of Missouri: trust department, Continental Illinois Bank and Trust Company, Chicago, June 1971.

Fred Obermiller, University of Missouri: minuteman

graduate education program, South Dakota State University, Sept. 1971.

James C. Ohls: assistant professor, department of economics, Princeton University, Sept. 1971.

James B. O'Neill: assistant professor, department of economics, University of Delaware, Sept. 1971.

Emmanuel O. Oyinola: research associate, department of economics, Iowa State University.

Neil A. Palomba: associate professor of economics, West Virginia University.

Richard A. Parker: instructor of economics, University of Illinois, Urbana.

Saroj Parasuraman: instructor, department of economics and business administration, State University of New York College at Fredonia, Aug. 1971.

Earl M. Peck, University of Colorado: assistant professor, department of economics, Western Illinois University.

Charles E. Phelps, University of Chicago: staff member, economics department, The Rand Corporation, Sept. 1971.

Alan A. Powell, Monash University, Australia: economist, Development Research Center, World Bank, Sept. 1971.

C. Wesley Randell, University of Missouri: foreign market development, American Soybean Association, Hudson Falls, Iowa, Aug. 1971.

C. Tait Ratcliffe: assistant professor, department of economics, Stanford University.

Fred H. Reuter, Virginia Polytechnic Institute and State University: assistant professor, department of economics, Clemson University.

R. Gene Reynolds: assistant professor of economics, Southwest Missouri State College, Sept. 1, 1971.

Yung W. Rhee, Johns Hopkins University: economist, Development Research Center, World Bank, Sept. 1971.

Edward A. Rice: assistant professor of accounting and business administration, Lebanon Valley College, Sept. 1971.

John A. Richards: assistant professor of economics, University of Hawaii.

Robert R. Richards: economist, National Bank of Alaska, Oct. 1971.

Sherman Robinson: assistant professor, department of economics, Princeton University, Sept. 1971.

Sheldon Rothstein: assistant professor of economics, Northeastern Illinois University, fall 1971.

Philip M. Scherer: assistant professor of economics, Virginia State College.

Joan D. Schneider: instructor, economics department, Middlebury College, Sept. 1971.

Bernard F. Sliger: professor of economics, Louisiana State University, Baton Rouge, Jan. 12, 1972.

Patrick J. Smith: acting assistant professor, department of economics; Graduate School of Administration; University of California, Riverside, July 1971.

William C. Smith: visiting instructor, department of economics, Mercer University, Sept. 1971.

William F. Staats, Federal Reserve Bank of Philadelphia: associate professor of finance, Louisiana State University, Baton Rouge.

William Staub, University of Missouri: economic

development branch, ERS, U.S. Department of Agriculture, Aug. 1971.

Stanley C. Stevens: instructor in economics, St. Olaf College, Sept. 1971.

Jacob A. Stockfisch, Institute for Defense Analyses: senior staff, economics department, The Rand Corporation, Aug. 1971.

Jai Myung Suh, University of Missouri: College of Business and Economics, Yonsu University, Seoul, Korea, Aug. 1971.

Ronald J. Sutherland, University of Oregon: assistant professor of economics, MacMurray College.

Jeannine Swift, State University of New York at Geneseo: assistant professor, Hofstra University.

Michael Szenberg: adjunct lecturer, department of economics, Brooklyn College of the City University of New York, Sept. 1, 1971.

Akira Takayama, Purdue University: visiting professor of economics, University of Hawaii, 1971-72.

Patricia D. Trainer, Queen Mary College, University of London: visiting assistant professor of economics, University of Illinois at Chicago Circle.

Alan T. Udall, Windham College: instructor, department of economics, University of Delaware, Sept. 1971.

Steve Van Meter, University of Missouri: loan applications department, Connecticut Mutual Life Insurance Company, Hartford, Dec. 1971.

Alejandro Velez, University of Florida: instructor, department of economics, Clemson University.

Darryl L. Webb: assistant professor of business law, University of Alabama.

John B. Weber: lecturer, department of economics and business administration, State University of New York College at Fredonia, Aug. 1971.

Claus Wittich, University of Southern California: economic affairs officer, Centrally Planned Economics Branch, ESA/CDPPP, United Nations, New York.

Louise B. Wolitz: instructor, department of economics, Fordham University, Sept. 1971.

Roland Y. Wu: assistant professor of economics, University of Maryland, Sept. 1, 1971.

Kozo Yamamura, Boston College: visiting professor of economics, University of Hawaii, 1971-72.

Baltasara Zalduendo: assistant professor, department of economics, Brooklyn College of the City University of New York, Sept. 1, 1971.

### *Leaves for Special Appointments*

Merrill J. Bateman, Brigham Young University: Mars Candy Co., London, Sept. 1, 1971.

Melvin Blase, University of Missouri-Columbia: Midwest Universities Consortium for International Activities, Michigan State University, Sept. 1971-Sept. 1972.

J. Hayden Boyd, Ohio State University: staff economist, program analysis division, Institute for Defense Analyses, Arlington, Virginia.

Hang-Sheng Cheng, Iowa State University: Office of the Deputy Assistant Secretary for International Affairs, U.S. Treasury, Washington.

Jack H. Clark, University of Guelph: Foundation Engineering Corporation of Canada, Pahang Tenggara Project, Sept. 1970-May 1972.

J. Malcolm Dowling, Jr., University of Colorado: Fulbright lectureship, University of Tehran, Iran, 1971-72.

Willard B. Doxey, Brigham Young University: University of Maryland, Germany, Sept. 1971.

John W. Hooper, University of California, San Diego: Ford Foundation faculty research fellowship, July 1, 1971-Mar. 31, 1972.

Louis Junker, Western Michigan University: UNESCO advisor to government of Mauritius, 1971-72.

Mordecai Kurz, Stanford University: Ford faculty fellowship.

William Mackenzie, McMaster University: Economic Research Bureau, Tanzania, 1971-72.

Elroy Mestre, State University College of New York at Oneonta: Organization of American States, Caracas, Venezuela, Sept. 1971-Sept. 1973.

Larry D. Neal, University of Illinois, Urbana: researcher, Organization for Economic Cooperation and Development, Paris, Sept. 1970-Aug. 1972.

Harry T. Oshima, University of Hawaii: visiting professor, University of the Philippines.

Douglas H. Pletsch, University of Guelph: lecturer in extension division, department of agricultural economics and farm management, University of Ghana, Sept. 1971.

John H. Power, University of Hawaii: visiting professor, Nairobi, Kenya.

Paul Y. Shin, California State College: Yonsei University, Seoul, Korea, 1971-72.

Arlon R. Tussing, University of Alaska: staff economist, U.S. Senate Committee on Interior and Insular Affairs.

Robert F. Wilcox, San Diego State College: head, Public Understanding of Science Office, Office of Government and Public Programs, National Science Foundation.

### *Resignations*

Mascell L. Beckford, University of Guelph: Canada Department of Finance, Ottawa, Aug. 15, 1971.

David E. Bond, University of British Columbia: executive director, Canadian Consumer Council, Ottawa, Jan. 1, 1972.

Robert Coen, Stanford University.

Claude Hillinger, Case Western Reserve University: Georgia Institute of Technology, June 1971.

Arnand P. Jaggi, State University of New York at Fredonia: Atlantic Christian College, Aug. 1971.

Barbara H. Kehrer, Fisk University.

Kenneth C. Kehrer, Fisk University.

Peter C. Mayer, Miami University: University of Guam, fall 1971.

V. Alonzo Metcalf, University of Missouri: Arizona State University, Sept. 1971.

Alan Milward, Stanford University.

Bridger Mitchell, Stanford University.

G. V. L. Narisimham, New York University: U.S. Department of Commerce.

Kenneth M. Parzych, State University of New York at Fredonia: Nichols College, Sept. 1971.

Melvin Reder, Stanford University.

Sarah H. Rodgers, University of Alabama.

George Rosen, Asian Development Bank, Philippines: New York Center for International Studies.

Marvin Rosenberg, Schools of Business, New York University.

Dean R. Sanders, Miami University, fall 1971.

#### *Miscellaneous*

William F. Butler, The Chase Manhattan Bank: chairman, New York State Council of Economic Advisers.

---

### NOTE TO DEPARTMENTAL SECRETARIES AND EXECUTIVE OFFICERS

When sending information to the *Review* for inclusion in the Notes Section, please use the following style:

A. Please use the following categories:

1—Deaths

2—Retirements

3—Foreign Scholars (visiting the USA or Canada)

4—Promotions

5—Administrative Appointments

6—New Appointments

7—Leaves for Special Appointments (NOT Sabbaticals)

8—Resignations

9—Miscellaneous

B. Please give the name of the individual (SMITH, John W.), his present place of employment or enrollment: his new title (if any), his next place of employment (if known or if changed), and the date at which the change will occur.

C. Type each item on a separate 3x5 card, and please do not send public relations releases.

D. The closing dates for each issue are as follows: *March*, November 1; *June*, February 1; *September*, May 1; *December*, August 1.

This announcement supersedes and replaces a letter which was sent annually from the managing editor's office.

---

## EMPLOYMENT SERVICES

### NATIONAL REGISTRY FOR ECONOMISTS

The National Registry for Economists was established in January, 1966, to provide a centralized nationwide clearinghouse for economists on a year-round basis. It is located in the Chicago Professional Placement Office of the Illinois State Employment Service and is staffed by experienced placement personnel, operating under the guidance and direction of Regional and National Bureau of Employment Security Professional Placement officials, and in cooperation with the American Economic Association. It is a free service. The National Registry for Economists maintains completely separate listings from those of the American Economic Association, and the National Registry does *not* have the listings as shown in the *American Economic Review*. There are no registration, referral, or placement fees. Application and order forms used in the Registry are available upon request from the: National Registry for Economists, Professional Placement Center, 208 South La Salle Street, Chicago, Illinois 60604.

### AMERICAN ECONOMIC ASSOCIATION VACANCIES AND APPLICATIONS

The Association renders this service to economists who wish to make known their availability for positions and to those who are seeking to fill vacancies in the field of economics. The Association takes no responsibility for selecting applicants or for following up the results. The Secretary's office will merely afford a central point for forwarding inquiries. The *Review* will publish a description of vacancies and applications submitted, with necessary editorial changes. The Secretary would appreciate receiving notification of appointments made as a result of this service. All inquiries about a listing with a key number should refer to it specifically. Communications should be addressed to: The Secretary, American Economic Association, 809 Oxford House, 1313 21st Avenue South, Nashville, Tennessee 37212.

Résumés and application blanks are *not* supplied by the American Economic Association. The Secretary's office will merely forward any announcements of positions or résumés received to a designated key number. Deadlines for the four issues of the *Review* are January 1, April 1, July 1, and October 1. Announcements will be repeated for a maximum of four consecutive issues unless a shorter period is requested by the advertiser.

#### *Vacancies*

*Economist:* Great teacher. Economics department in well known and highly desirable college town seeks an outstanding undergraduate teacher. Must have Ph.D. and publications appropriate for experience. Courses containing moderate numbers of students. Our department competes for undergraduate majors with other departments containing nationally recognized scholars, teachers, Pulitzer Prize winners, etc. Salary and rank are open. Please do not apply unless it is possible that your references will classify you as a truly magnificent teacher whose work motivates and inspires students. Initial visiting appointments can be arranged. Note: faculty and students

are requested to nominate candidates for this position from among the great teachers whom they have known. P371

*Industry Economist:* Operations Group of the Federal Communications Commission, responsible for policy matters, legal and administrative activities of the Spectrum Management Task Force, has vacancy for economist to analyze and investigate the varied and technical economic factors pertaining to spectrum management. Applicant must be familiar with the application of computers to data analysis in conducting specific studies. Position will be filled at GS-11, 12 or 13

# The American Economic Review

## CONTENTS

1. **THE EFFECTS OF THE 1970-71 RECESSION ON THE U.S. ECONOMY: A REVIEW OF THE EVIDENCE** ..... 1  
 by **JOHN F. HANSEN** ..... 1

2. **THE EFFECTS OF THE 1970-71 RECESSION ON THE U.S. ECONOMY: A REVIEW OF THE EVIDENCE** ..... 1  
 by **JOHN F. HANSEN** ..... 1

3. **THE EFFECTS OF THE 1970-71 RECESSION ON THE U.S. ECONOMY: A REVIEW OF THE EVIDENCE** ..... 1

4. **THE EFFECTS OF THE 1970-71 RECESSION ON THE U.S. ECONOMY: A REVIEW OF THE EVIDENCE** ..... 1  
 by **JOHN F. HANSEN** ..... 1

5. **THE EFFECTS OF THE 1970-71 RECESSION ON THE U.S. ECONOMY: A REVIEW OF THE EVIDENCE** ..... 1  
 by **JOHN F. HANSEN** ..... 1

6. **THE EFFECTS OF THE 1970-71 RECESSION ON THE U.S. ECONOMY: A REVIEW OF THE EVIDENCE** ..... 1

7. **THE EFFECTS OF THE 1970-71 RECESSION ON THE U.S. ECONOMY: A REVIEW OF THE EVIDENCE** ..... 1  
 by **JOHN F. HANSEN** ..... 1

8. **THE EFFECTS OF THE 1970-71 RECESSION ON THE U.S. ECONOMY: A REVIEW OF THE EVIDENCE** ..... 1  
 by **JOHN F. HANSEN** ..... 1

9. **THE EFFECTS OF THE 1970-71 RECESSION ON THE U.S. ECONOMY: A REVIEW OF THE EVIDENCE** ..... 1

10. **THE EFFECTS OF THE 1970-71 RECESSION ON THE U.S. ECONOMY: A REVIEW OF THE EVIDENCE** ..... 1  
 by **JOHN F. HANSEN** ..... 1

THE AMERICAN ECONOMIC REVIEW  
 PUBLISHED BY THE AMERICAN ECONOMIC ASSOCIATION

# THE AMERICAN ECONOMIC ASSOCIATION

Founded in 1885

• Published at George Banta Co., Inc., Menasha, Wisconsin.

• THE AMERICAN ECONOMIC REVIEW, including four quarterly numbers, the *Proceedings* of the annual meetings, and *Directory* and Supplements, is published by the American Economic Association and is sent to all members five times a year, in March, May, June, September, and December.

• Membership dues of the Association are \$20.00 a year, which includes a year's subscription to both the *American Economic Review* and the *Journal of Economic Literature*. Subscriptions by nonmembers are \$30.00 a year, and only subscriptions to both publications will be accepted. Single copies of the *Review* and *Journal* are \$4.00 each. Each order for copies of either publication must also include a \$.50 per order service charge. Orders should be sent to the Secretary's office, Nashville, Tennessee.

• Correspondence relating to the *Papers and Proceedings*, the *Directory*, advertising, permission to quote, business matters, subscriptions, membership and changes of address may be sent to the secretary, Rendigs Fels, 1313 21st Avenue, South, Nashville, Tennessee 37212. To be effective, notice of change of address must reach the secretary by the 1st of the month previous to the month of publication. The Association's publications are mailed by second class and are not forwardable by the Post Office.

• Second-class postage paid at Nashville, Tennessee and at additional mailing offices. Printed in U.S.A.

## Officers

### *President*

JOHN KENNETH GALBRAITH  
Harvard University

### *President-Elect*

KENNETH ARROW  
Harvard University

### *Vice-Presidents*

HENDRIK S. HOUTHAKKER  
Harvard University  
ARTHUR M. OKUN  
Brookings Institution

### *Secretary-Treasurer and Editor of Proceedings*

RENDIGS FELS  
Vanderbilt University

### *Managing Editor of The American Economic Review*

GEORGE H. BORTS  
Brown University

### *Managing Editor of The Journal of Economic Literature*

MARK PERLMAN  
University of Pittsburgh

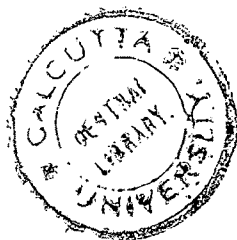
## Executive Committee

### *Elected Members of the Executive Committee*

ROBERT DORFMAN  
Harvard University  
ARNOLD C. HARBERGER  
University of Chicago  
ROBERT EISNER  
Northwestern University  
JOHN R. MEYER  
Yale University  
GUY HENDERSON ORCUTT  
Yale University  
JOSEPH A. PECHMAN  
Brookings Institution

### *Ex Officio Members*

WASSILY LEONTIEF  
Harvard University  
JAMES TOBIN  
Yale University



## NICHOLAS GEORGESCU-ROEGEN

DISTINGUISHED FELLOW

1971

We honor Nicholas Georgescu-Roegen, whom Paul Samuelson has called "a scholar's scholar and an economist's economist." Ever since the publication of his path-breaking paper on the Pure Theory of Consumer's Behavior 35 years ago, Georgescu-Roegen has been a seminal writer on utility theory, in which almost every important theoretical problem or relevant issue bears the mark of his pen. In production theory, where he has been one of the pioneers of the theory of linear systems, his contributions extend from the two most important theorems in the input-output analysis to a radical reformulation of the concept of production function. No American economist has more successfully combined in his training and publications the fields of economics, mathematics, and statistics. Yet Georgescu-Roegen has remained a signal defender of the view that many important problems are beyond the reach of numbers. Unique also is his keen knowledge of past and present human institutions. To this knowledge, which ranks very high in his intellectual hierarchy, we owe his penetrating contributions to institutional economics with their highly original adaptation of analytical tools to a complex structure. A recent product of his broad range of knowledge is his new book, *The Entropy Law and the Economic Process*, in which he develops the revolutionary view that economic activity is an extension of man's biological evolution—an entropic process rather than the mechanical analogue traditional in mathematical economics. At the same time, Georgescu-Roegen has been and still is a great teacher. Every one of his class lectures is a work of a long-mastered art, the pride of all those who were fortunate to be his students and to know his continuous and friendly devotion to their academic progress and career. To Nicholas Georgescu-Roegen—the scholar, the teacher, the humanist—we pay homage as a true Renaissance man.



*Nicholas Georgescu-Roegen*

# Maximum Principles in Analytical Economics

By PAUL A. SAMUELSON\*

The very name of my subject, economics, suggests economizing or maximizing. But Political Economy has gone a long way beyond home economics. Indeed, it is only in the last third of the century, within my own lifetime as a scholar, that economic theory has had many pretensions to being *itself* useful to the practical businessman or bureaucrat. I seem to recall that a great economist of the last generation, A. C. Pigou of Cambridge University, once asked the rhetorical question, "Who would ever think of employing an economist to run a brewery?" Well, today, under the guise of operational research and managerial economics, the fanciest of our economic tools are being utilized in enterprises both public and private.

So at the very foundations of our subject maximization is involved. My old teacher, Joseph Schumpeter, went much farther. Instead of being content to say economics must borrow from logic and rational empirical enquiry, Schumpeter made the remarkable claim that man's ability to operate as a logical animal capable of systematic empirical induction was itself the direct outcome of the Darwinian struggle

for survival. Just as man's thumb evolved in the struggle to make a living—to meet his economic problem—so did man's brain evolve in response to the economic problem. Coming forty years before the latest findings in ethology by Konrad Lorenz and Nikolaas Tinbergen, this is a rather remarkable insight. It would take me away from my present subject to more than mention the further view enunciated by Schumpeter [42] in launching the new subject of econometrics. Quantity, he said, is studied by the physicist or other natural scientist at a fairly late and sophisticated stage of the subject. Since a quantitative approach is, so to speak, at the discretion of the investigator, all the more credit to the followers of Galileo and Newton for taking the mathematical approach. But in economics, said Schumpeter, the very subject matter presents itself in quantitative form: take away the numerical magnitude of price or barter exchange-ratio and you have nothing left. Accounting does not benefit from arithmetic; it is arithmetic—and in its early stages, according to Schumpeter, arithmetic is accounting, just as geometry in its early stages is surveying.

I must not leave you with the impression that analytical economics is concerned with maximization principles primarily in connection with providing vocational handbooks for the practicing decision maker. Even back in the last generation, before economics had pretensions toward being itself useful to practitioners, we economists were occupied with maxima and minima. Alfred Marshall's *Principles of Economics*, the dominating treatise in the forty years

\* Professor of economics, Massachusetts Institute of Technology; 1970 Nobel Memorial Laureate of Economic Science. This article is the lecture he delivered in Stockholm, Sweden, December 11, 1970, when he received the Nobel Prize in Economic Science. Minor corrections and additions have been made by the author. The article is published here with the permission of the Nobel Foundation and is included in the complete volume of *Les Prix Nobel en 1970* as well as in the series Nobel Lectures (in English) published by the Elsevier Publishing Company, Amsterdam and New York. The article was also printed in *Science*, September 10, 1971, 173, 991-97.

after 1890, dealt much with optimal output at the point of maximum net profit. And long before Marshall, A. A. Cournot's 1838 classic, *Researches into the Mathematical Principles of the Theory of Wealth*, put the differential calculus to work in the study of maximum-profit output. Concern for minimization of cost goes back a good deal more than a century, at least back to the marginal productivity notions of von Thünen.

It is fashionable these days to speak of identity crises. One must not make the mistake attributed to Edward Gibbon when he wrote his *Decline and Fall of the Roman Empire*. Gibbon, it was said, sometimes confused himself and the Roman Empire. I know in these days of the living theater—and I ought to add on this occasion, of the theories of quantum mechanics—the distinction often becomes blurred between observing audience and acting players, between the observing scientist and the guinea pigs or atoms under observation. As I shall discuss in connection with the role of maximum principles in natural science, the plumb-line trajectory of a falling apple and the elliptical orbit of a wandering planet may be capable of being described by the optimizing solution for a specifiable programming problem. But no one will be tempted to fall into a reverse version of the Pathetic Fallacy and attribute to the apple or the planet freedom of choice and consciously deliberative minimizing. Nonetheless, to say "Galileo's ball rolls down the inclined plane *as if* to minimize the integral of action, or to minimize Hamilton's integral," does prove to be useful to the observing physicists, eager to formulate predictable uniformities of nature.

What is it that the scientist finds useful in being able to relate a positive description of behavior to the solution of a maximizing problem? That is what a good deal of my own early work was about. From the

time of my first papers on "Revealed Preference" [21], [22], [25], [29], through the completion of *Foundations of Economic Analysis*, I found this a fascinating subject. The scientist, as with the housewife, finds his work is really never done. Just in these last weeks I have been working on the very difficult problem of understanding stochastic speculative price—e.g., how cocoa prices fluctuate on the London and New York exchanges [39]. When confronted with an unmanageable system of non-linear difference equalities and inequalities, I could have despaired of finding in the mathematical literature a proof of even the existence of a solution. But suddenly the problem became solvable in a flash, when out of the strata of memory, I dredged up the recollection that my positive descriptive relations could be interpreted as the necessary and sufficient conditions of a well-defined maximum problem. But I run ahead of my story if I give you the impression that maximum principles are valuable merely as a convenience and crutch to the less-than-omniscient analyst.

Seventy years ago, when the Nobel Foundation was first established, the methodological views of Ernst Mach enjoyed a popularity they no longer possess.<sup>1</sup> Mach you will remember, said that what the scientist seeks is an "economical" description of nature. By this he did not mean that the navigation needs of traders decreed that Newton's system of the world had to get born. He meant rather that a good explanation is a simple one that is easy to remember and one which fits a great variety of the observable facts. It would be a Gibbonlike fallacy to illustrate this by the deistic view of Maupertuis that

<sup>1</sup> Whatever their ultimate worth, we must be grateful to Mach's concepts for their influence on the young Einstein's formulation of special relativity theory. Although an older Einstein rebelled against this same methodology, this cannot rob Mach's concepts of their just credit.

the laws of nature are the working out of a simple teleological purpose. Mach is not saying that Mother Nature is an economist; what he is saying is that the scientist who formulates laws of observed empirical phenomena is essentially an economist or economizer.

Nonetheless, I must point out that these distinct roles are almost by coincidence so to speak, closely related. Often the physicist gets a better, a more economical, description of nature if he is able to formulate the observed laws by a maximum principle. Often the economist is able to get a better, more economical, description of economic behavior from the same device.

Let me illustrate this by some very simple examples. Newton's falling apple can be described in either of two ways: its acceleration toward the earth is a constant; or its position as a function of time follows that arc which minimizes the integral, taken from its moment of release to the terminal time at which it is observed, of an integrand which can be written as the square of its instantaneous velocity minus a linear function of its position. "What?" you will say, "can you seriously regard the second explanation as the simple one?" I will not argue the point, except to point out that simplicity is in the eye of the beholder, and that if I were to write out for the mathematical physicist the expression

$$\delta \int_0^T \left( \frac{1}{2} \dot{x}^2 - gx \right) dt = 0$$

he would not consider it less simple than  $\ddot{x} = -g$ ; and he would know that the Hamilton principle formulation in variational terms has great mnemonic properties when it comes to transforming from one coordinate system to another.

Although I am not a physicist and do not suppose that many of my audience are either, let me give a clearer example of the usefulness of a minimum principle in physics. Light travels between two points

in the air before me along a straight line. Alternatively like the apple's fall, this arc can be defined as the solution to a minimization problem in the calculus of variations. But now let us consider how light is reflected when it hits a mirror. You may observe and memorize the rule that the angle of reflection is equal to the angle of incidence. A neater way of understanding this fact is by the least-time principle of Fermat, which was already known to Hero and other Greek scientists. The accompanying diagram with its indicated similar triangles can be self-explanatory.

If  $ABC'$  is clearly shorter in length than  $ADC'$ , it is evident that the similar  $ABC$  path is shorter and involves less time than any other path such as  $ADC$ .

You could validly argue that the minimum formulation is neat, but really no better than the other formulation. However, move from this lecture room to your bathtub and observe your big toe in the water. Your limbs no longer appear straight because the velocity of light in water differs from that in air. The least-time principle tells you how to formulate behavior under such conditions and the memorizing of Snell's Law about angles does not. Who can doubt which is the better scientific explanation?

#### An Illustrative Economic Example

Let me illustrate the same thing in economics as a simplest imaginable case. Consider a profit-maximizing firm that sells its output along a demand curve in which the price received is a nonincreasing function of the amount sold. Suppose further that output is producible by two, three, or ninety-nine different inputs. To keep the example simple, suppose that the production function relating outputs to inputs is smooth and concave.

As a positivistic scientist interested merely in cataloguing the observable facts a Machian economist could in principle

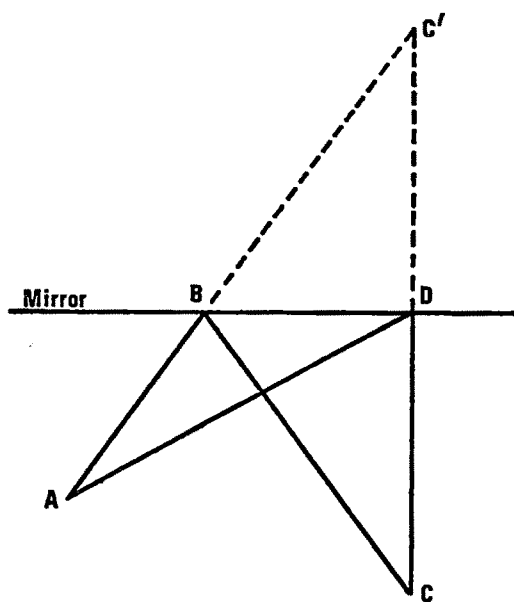


FIGURE 1

record on punch cards ninety-nine demand functions relating the quantity of each input bought by the firm to the ninety-nine variables depicting the input prices. What a colossal task it would be to store bits of information defining ninety-nine distinct surfaces in a one-hundred dimensional space! But the ninety-nine surfaces are not really independent. In actuality, it is enough to have knowledge of a single parent surface in order to be able to calculate the exact information about the ninety-nine children. How is this tremendous economy of description possible? It is by virtue of the fact that the observed demand curves, which that great Swedish economist of the generation before last, Gustav Cassel, would have taken as the irreducible atoms of the economists' theory, are actually themselves solutions to a maximum-profit problem. Under simple regularity conditions of the calculus, they are the inverse functions of a family of partial derivatives of the Total Revenue function, where revenue is given by the output producible by any specified quanti-

ties of all inputs times the determinate demand price at which that output will sell. When smooth and strongly concave, this parent revenue function has as its children a ninety-nine by ninety-nine matrix of second partial derivatives which is symmetric and negative definite. It is an exercise in algebra to show that these functions can be uniquely inverted to form a new family of children with the same properties; and ninety-nine such children cannot fail to have a parent function which, so to speak, if it had never existed we should have to invent in a Pygmalion fashion.

Mathematically we have

$$\begin{aligned} & \text{Max}_{\{v_i\}} \left[ R(v_1, \dots, v_{99}) - \sum_1^{99} p_j v_j \right] \\ (1) \quad & = R(v_1^*, \dots, v_{99}^*) - \sum_1^{99} p_j v_j^* \\ & = -H(p_1, \dots, p_{99}), \end{aligned}$$

where  $R(v_1, \dots, v_{99})$

$$= Q(v_1, \dots, v_{99})P[Q(v_1, \dots, v_{99})]$$

is a smooth, strongly concave "regular" revenue function. Necessary conditions for the maximum are

$$\begin{aligned} & \partial R(v_1^*, \dots, v_{99}^*) / \partial v_i = p_i, \\ (2) \quad & (i = 1, \dots, 99) \end{aligned}$$

If in addition the Hessian matrix of second partial derivatives,  $[\partial^2 R / \partial v_i \partial v_j]$ , is negative definite, equations (2) are sufficient for the maximum. This implies inverse relations that can be interpreted as partial derivatives of a Hotelling-Roy dual function,  $H$ , namely

$$\begin{aligned} (3) \quad & v_i^* = \partial H(p_1, \dots, p_n) / \partial p_i, \\ & (i = 1, \dots, 99) \end{aligned}$$

It follows that for

$$\sum \Delta v_j^2 \neq 0 \neq \sum \Delta p_j^2,$$

our variables satisfy the inequality

$$(4) \quad \Delta p_1 \Delta v_1 + \Delta p_2 \Delta v_2 + \dots + \Delta p_{99} \Delta v_{99} < 0$$

More can be said. Although my intuition is poor enough in three dimensional space, I can assert with confidence on the basis of the above that raising any input's price while holding all remaining inputs' prices constant will definitely reduce the amount demanded of that input by the firm—i.e.,  $\partial v_i / \partial p_i < 0$ : Such a commonsense result might be expected by anyone who performed an act of emphathetic introspection, "Suppose I were a jack-ass of an entrepreneur, what would I do to adjust to the dearness of an input in order to conserve as much profit as possible?"

Here the commonsense and advanced mathematics happen to agree. But we all know the Giffen pathology according to which an increase in the price of potatoes to Irish peasants, who must depend heavily on potatoes when they are poor, may itself impoverish them so as to force them into buying more rather than less potatoes. In this case common sense recognizes itself only under the search-light of mathematics.

With the assistance of mathematics, I can see a property of the ninety-nine dimensional surfaces hidden from the naked eye. If an increase in the price of fertilizer alone always increases the amount the firm buys of caviar, from that fact alone I can predict the answer to the following experiment which I have never seen performed and upon which I have no observations: an increase in the price of caviar alone will increase the amount the firm buys of fertilizer. In thermodynamics such reciprocity or integrability conditions are known as Maxwell's Conditions; in economics they are known as Hotelling conditions in honor of Harold Hotelling's 1932 work [9].

One of the pleasing things about science is that we do all climb towards the heavens on the shoulders of our predecessors. Economics, like physics has its heroes, and the letter *H* that I used in my mathematical equations was not there to honor Sir Wil-

liam Hamilton, but rather Harold Hotelling. For it was his work that I found so stimulating when I came on the scene, at about the same time that the late Henry Schultz [41] was trying by econometric methods to verify the empirical validity of the Hotelling integrability conditions.

There are still other predictable conditions of definiteness relating to how weak these "cross effects" must be in comparison with "own effects," but I will spare my audience discussion of them, except for mention of the condition that all principal minors have to oscillate in sign.

As a last illustration of the black magic by which a maximum formulation permits one to make clearcut inferences about a complicated system involving a large number of variables, let me recall the work I have done in formulating clearly and generalizing what is known in physics as LeChatelier's Principle [24], [33], [32]. This Principle was enunciated almost one hundred years ago by a French physicist interested in Gibbs-like thermodynamics. It is a vague principle. A third of a century ago when I thumbed through different physics treatises, my mathematical ear could not discern what tune was being played. If you pick up most physics books today, perhaps your luck will be no better. Usually the argument is obscurely teleological, reading something like the following: If you put an external constraint on an equilibrium system, the equilibrium shifts to "absorb" or "resist" or "adjust to" or "minimize" the change. I was struck by a remark made by an old teacher of mine at Harvard, Edwin Bidwell Wilson. Wilson was the last student of J. Willard Gibbs at Yale and had worked creatively in many fields of mathematics and physics: his advanced calculus was a standard text for decades; his was the definitive writeup of Gibbs' lectures on vectors; he wrote one of the earliest texts on aerodynamics; he was a friend of R. A. Fisher and an expert on

mathematical statistics and demography; finally, he had become interested early in the work of Pareto and gave lectures in mathematical economics at Harvard. My earlier formulation of the inequality in equation (4) owed much to Wilson's lectures on thermodynamics. In particular I was struck by his statement that the fact that an increase in pressure is accompanied by a decrease in volume is not so much a theorem about a thermodynamic equilibrium system as it is a mathematical theorem about surfaces that are concave from below or about negative definite quadratic forms. Armed with this clue I set out to make sense of the LeChatelier Principle.

Let me now enunciate a valid formulation of that principle. "Squeeze a balloon and its volume will contract. But compare how its volume contracts under two different experimental conditions. First, imagine that its surface is insulated from the rest of the world so that none of the so-called heat engendered can escape. In the second alternative administer the same increase in pressure in the balloon, but let it come into temperature equilibrium with the unchanged temperature of the room. Then according to LeChatelier Principle the decrease in volume when the insulation constraint is placed on the system will be *less* than when the temperature is constrained to end up constant." The steeper light curve in Figure 2 shows the relationship between the pressure on the vertical axis and volume on the horizontal axis that prevails for the insulated increase. The less steep curve going through the same point *A* shows the pressure-volume relationship for an iso-thermal change. It is the essence of LeChatelier's Principle that the light curve must be more steep than the heavy curve, or, in usual thermodynamic notation.

$$(5) \quad (\partial v / \partial p)_t \leq (\partial v / \partial p)_s \leq 0$$

where *t* stands for temperature held con-

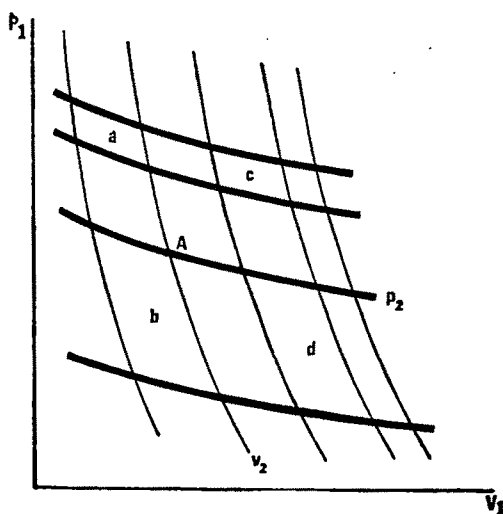


FIGURE 2

stant, and *s* stands for the insulated (or adiabatic or isentropic) change.

Now what in the world has all this to do with economics? There is really nothing more pathetic than to have an economist or a retired engineer try to force analogies between the concepts of physics and the concepts of economics. How many dreary papers have I had to referee in which the author is looking for something that corresponds to entropy or to one or another form of energy. Nonsensical laws, such as the law of conservation of purchasing power, represent spurious social science imitations of the important physical law of the conservation of energy; and when an economist makes reference to a Heisenberg Principle of indeterminacy in the social world, at best this must be regarded as a figure of speech or a play on words, rather than a valid application of the relations of quantum mechanics.

However, if you look upon the monopolistic firm hiring ninety-nine inputs as an example of a maximum system, you can connect up its structural relations with those that prevail for an entropy-maximizing thermodynamic system. Pressure and volume, and for that matter absolute tem-

perature and entropy, have to each other the same conjugate or dualistic relation that the wage rate has to labor or the land rent has to acres of land. Figure 2 can now do double duty, depicting the economic relationships as well as the thermodynamic ones. Now on the vertical axis goes  $p_1$ , the price of the first input. On the horizontal axis goes  $v_1$ , its quantity. The story can be told of a ninety-nine variable system but I think you will forgive me if I discuss the simpler case of two variables, say labor and land.

As in the case of the balloon we perform an experiment under two alternative sets of specified conditions. In the first case, we raise  $p_1$ , the price of the first input labor, while holding constant the quantity of the second input, land or  $v_2$ —as for example in the Marshallian short run when only labor can be varied. The rise in  $p_1$  must lower  $v_1$  as shown by the negative slope of the light curve through  $A$ .

Now, in the second alternative experiment, we raise  $p_1$  by the same amount but hold the price of  $v_2$ ,  $p_2$ , constant. Again, for a profit maximizing monopolist there can be but one qualitative answer: less of  $v_1$  will now be bought, as shown by the negative slope of the heavy curve through  $A$ . Now one can state what perhaps might be called the LeChatelier-Samuelson Principle: The heavy curve of longer-run adjustment, with other price constant (and other quantity of course thereby itself adjusting *mutatis mutandis* to restore the maximum-profit equilibrium), must be less steep or more elastic than the light curve depicting the demand reaction when the other input is held constant. Mathematically now

$$(6) \quad (\partial v_1 / \partial p_1)_{p_2} \leq (\partial v_1 / \partial p_1)_{v_2} \leq 0$$

I have included the equality signs to allow for the case where two inputs might be quite independent in production. What is remarkable about the relation is that the

indicated inequalities will hold whether the two inputs are complements such as pumps and insecticides or substitutes such as organic and inorganic fertilizers. The interested listener might try to work out the intuitive verification of this in those opposite cases.

Not only in the theory of production but also in the general theory of constrained rationing does the LeChatelier Principle have various economic applications.

### Consumer Demand Theory

This brings me to the theory of consumer demand. Unlike the maximizing profit situation that has been discussed up to this time, now we have a budgetary constraint within which maximizing has to occur. Prior to the mid-1930's, utility theory showed signs of degenerating into a sterile tautology. Psychic utility or satisfaction could scarcely be defined, let alone be measured. Austrian economists would insist that people acted to maximize their utility, but when challenged as to what that was, they found themselves replying circularly that however people behaved, they would presumably not have done so unless it maximized their satisfaction. Just as we can cancel two from the ratio of even numbers, so one could use Occam's Razor to cut utility completely from the argument, ending one up with the fatuity: people do what they do.

I exaggerate only a little. It is true that the Russian Slutsky [43] had in 1915 gone beyond this, but his work, published in an Italian journal, was forgotten in the backwash of the First World War. The better known work of Pareto [16] [17] lacked the mathematical technique of the Weierstrass theory of constrained extrema. Two dimensional analysis of indifference curves had been worked out by W. E. Johnson [12], a Cambridge logician who had studied with Marshall and Whitehead and who is thought to have influenced the probabil-

ity researches of J. M. Keynes [13], Frank Ramsey [20], and Sir Harold Jeffreys [11]. However just before I arrived on the scene, when Sir Roy Allen and Sir John Hicks [8] at the London School, and Henry Schultz in Chicago, were pioneering the theory of consumers' behavior, the contributions of Slutsky were unknown.

From the beginning I was concerned to find out what *refutable* hypotheses on the observable facts on price and quantity demanded were implied by the assumption that the consumer spends his limited income at given prices in order to maximize his ordinal utility (i.e., his better-or-worse situation without regard to any numerical indication of how much better or worse). To make a long story short, the flash of inspiration for "Revealed Preference" came to me in argument with one of my teachers, as so many of my best ideas have done. Having learned about indifference curves from Leontief, I put them to use next year in Haberler's international trade course. When he objected to my postulating convex indifference curves, I heard myself replying: "Well, if they are concave, then the Laspeyres-Paasche index-numbers of your doctoral thesis are no good."<sup>2</sup> Far from being a *reductio ad absurdum*, this proposition, upon reflection, suggested how a scientific investigator could refute the hypothesis of maximizing behavior by a test on two price-quantity observed situations. All that remained was to work out the details of the theory of revealed preference.

My early theory of revealed preference was by itself perfectly adequate to handle

the problems of two consumption goods. I went on to conjecture that if we ruled out similar contradictions for choices of more than two situations,<sup>3</sup> then the phenomenon of "nonintegrability" of the indifference field could be ruled out.

Especially on occasions like this when one is only too likely to reminiscence about scientific victories, one ought to pause frequently along the way to express some lamentations over defeats and failures. Even with the aid of some of the world's leading mathematicians I was not able to verify and prove the truth of the previous footnote's conjecture, and I was persuaded to omit that material from the published version of "Revealed Preference." All the more credit therefore must go to Hendrik Houthakker [10] who on his maiden venture into economics formulated the Strong Axiom and proved that it did exclude nonintegrability.

How shall I in a morning lecture explain in words what nonintegrable indifference fields are all about? In 1950 I [28] gave a review of the integrability discussion, going back to Pareto in the early years of this century and before that to Irving Fisher's 1892 classic thesis [5], and even before that to resurrected work of the rather unknown Antonelli [1]. How obscure the status of the integrability problem was in the mid-1930's when I arrived on the scene can be indicated by the fact that two close collaborators already cited, Sir John Hicks and Sir Roy Allen, seemed actually to be at odds in their views on the subject. Now that the empirical implications of nonintegrability are understood, most theorists are inclined to postulate integrability. How to make clear its meanings? My good friend

<sup>2</sup> In explanation, suppose you are maximizing the utility of your consumptions of  $(Q_x, Q_y, \dots)$ , at prices  $(P_x, P_y, \dots)$ , spending positive income of  $P_x Q_x + \dots = \sum P Q$ . Then in two situations,  $(P^1, Q^1, \sum P^1 Q^1)$  and  $(P^2, Q^2, \sum P^2 Q^2)$ , it is a contradiction to maximization of ordinal utility to be able to observe both  $\sum P^1 Q^2 / \sum P^1 Q^1 < 1$ , and  $\sum P^2 Q^1 / \sum P^2 Q^2 < 1$ . With variants of  $\leq$  for  $<$ , denying this possibility is one form of the Weak Axiom of Revealed Preference.

<sup>3</sup> Using the notation of the previous situation, I conjectured that nonintegrability could be ruled out by the axiom " $\sum P^i Q^i > \sum P^i Q^{i+1}$  for all  $i=1, \dots, n-1 \geq 1$  rules out  $\sum P^n Q^n > \sum P^n Q^1$ ." For  $n=2$ , this merely repeats the Weak Axiom; for all  $n \geq 2$ , it becomes the Strong Axiom of Houthakker.

Nicholas Georgescu-Roegen [6], from whose classic 1936 paper I gleaned so many insights into the integrability problem, would argue that it is impossible to state such complicated mathematical relations in mere words. I am on record with the contrary view, namely that mathematics is language and in principle what one fool can comprehend so can another. Let me therefore refer you to Figure 2 in which I am able to present an in-the-large interpretation of integrability conditions for our earlier profit maximizing firm with its hiring of ninety-nine inputs.

The steep curves in the diagram represent the demand functions for the first input  $v_1$ , in terms of its price  $p_1$ , when all other inputs are held constrained as in a Marshallian short run. The heavier and less steep curves also represent demand functions for  $v_1$  in terms of  $p_1$ , but with all other factor prices frozen. If someone challenged me to explain what the existence of integrability implies, but refused to let me use the language of partial derivatives, I could illustrate by an equi-proportional-area property in Figure 2 that meaning of integrability. I may say that the idea for this proposition in economics came to me in connection with some amateurish researches in the field of thermodynamics. While reading Clerk Maxwell's charming introduction to thermodynamics, I [35] found that his explanation of the existence of the same absolute temperature scale in every body could be true only if on the  $p$ - $v$  diagram that I earlier referred to in connection with LeChatelier's Principle, the two families of curves—steep and light or less-steep and heavy—formed parallelograms like  $a$ ,  $b$ ,  $c$ ,  $d$  in Figure 2 which everywhere have the property  $\text{area } a / \text{area } b = \text{area } c / \text{area } d$ . And so it is with the two different economic curves. It is a consequence of the Hotelling integrability conditions which link together the ninety-nine different demand functions for factors

that the areas shown have this proportionality property. In leaving this interesting result, let me mention that it holds even when—as in linear programming—the relevant surfaces have corners and edges along which unique partial derivatives are not defined. Finally, this illustrates that once we know one of the demand functions everywhere, the other function only needs to be known along one razor's edge in space in order for it to be determined everywhere.

I should not leave the analytics of maximizing functions without mentioning that all of this is not an idle exercise in logic and mathematics.<sup>4</sup> Debates rage in economics

<sup>4</sup> In reaction to the previously published version of this lecture, Robert Killingsworth, a graduate student at Yale, has written to point out that often in physics no nice distinction is made between a maximum and a minimum, or for that matter between an extremum of either kind and a stationary inflection point. I quite agree, and have often had occasion to point out that the physicist typically makes use only of the "variational" aspect of the problem—as for example in my paper on causality and teleology in economics in D. Lerner, ed., *Cause and Effect* (New York 1965, pp. 99–143, particularly p. 128). Thus, I can throw a ball to bean you in two ways: by direct fire, or by throwing it so high that it falls on you in indirect fire. The first trajectory does, and the second does not, truly minimize the "action" integral. So to speak, just as Nature abhors a vacuum only up to 30 inches of mercury, she is a myopic minimizer who only minimizes action up to the first conjugate point ahead. In still other contexts, as for example my path-of-light case, the physicist does not really believe that there is anything teleological going on: he thinks of light waves as going off from every point in all directions, in accordance with Huyghen's principle; and he expects that such waves will reinforce and counteract each other at various points; what is seen as a ray of light in geometric optics is simply those loci where the amount of cancellation of waves has been least. In terms of economics, this is rather like the Darwinian arguments of Armen Alchian of 1950 in the *Journal of Political Economy*, namely that the survival of the fittest presents us with phenomena that act as if they came from an extremum problem. This has the following consequence, as Killingsworth has pointed out in referring to A. d'Abro, *Decline of Mechanism* (Princeton 1939, ch. 18): If in my first figure, we bend the mirror around  $B$ , preserving its slope there but making it more curved than the ellipse defined by  $A$  and  $C$  as foci, then the actual path that light takes as seen from  $A$  to  $B$  to  $C$  will have the *longest*, rather than the shortest distance! In still other cases, the actual path might

as to whether corporations maximize their profits. Yet neither side of the debate pauses to ask what difference it ought to make for observables if there is or there is not some function that is being maximized. And if I depart from the narrow field of economics, I must confess that the writings of sociologists like Talcott Parsons [18] seem to me to be seriously empty because they never seem even to ask the question of what difference it makes to have social action part of a maximizing value system, or just what is implied by "functionalist" interpretations of the observed phenomena.

### Nonmaximum Problems

I must not be too imperialistic in making claims for the applicability of maximum principles in theoretical economics. There are plenty of areas in which they simply do not apply. Take for example my early paper dealing with the interactions of the accelerator and the multiplier [23]. This is an important topic in macroeconomic analysis. Indeed, as I have recorded elsewhere this paper brought me a disproportionate amount of reputation. True the topic was a fundamental one, and mathematical analysis of stability conditions was able to give it a neat solution at a level that could be understood both by the intelligent beginner and the virtuoso in mathematical economics. But the original specification of the model had been made by my Harvard teacher Alvin Hansen, and the works of Sir Roy Harrod [7] and Erik Lundberg

be made to seem to be neither a maximum nor a minimum, merely a stationary inflection point (or, loosely speaking, a saddlepoint). With some straining, one can fit this in under the case of a conjugate point, as above, by the following consideration: Simultaneously halve and re-halve the distances of *A* and *C* from *B*, until eventually you will be able to say that the finite path that the light takes is indeed a true minimum; or, to generalize, in geometrical optics, for *sufficiently near-together points* on the path that light "follows," the proper Hero-Fermat-Maupertuis integral is truly *minimized*. It must be stressed that in economics, true minimization is what is important because the actors are postulated to have purpose from the beginning.

[15] clearly pointed the way to the setting up of this model.

My point in bringing up the accelerator-multiplier here is that it provides a typical example of a dynamic system that can in no useful sense be related to a maximum problem. By examining the sick we learn something about those who are well; and by examining those who are well we may also learn something about the sick. The fact that the accelerator-multiplier *cannot* be related to maximizing takes its toll in terms of the intractability of the analysis. Thus when my colleague, Professor Richard Eckaus, was a younger man, he wrote a doctoral dissertation [4] under my direction on generalizing the accelerator-multiplier analysis to many sectors and countries. It was an excellent piece of scholarship; Dr. Eckaus, with great ingenuity and elegance, extracted everything from the model that could be extracted. Yet he would be the first to assert that, in a sense, the ratio of useful output to high grade input was somewhat disappointing. Few grand simplicities emerged. The conscientious investigator had to point out a great range of possibilities that could happen, and had to use up much of his intelligence in taxonomy and classification of those possibilities. To illustrate the intrinsic intractability of such a problem, let me recall to you a remarkable difficulty. Suppose Europe in 1970 is a seventeen sector multiplier-accelerator complex that is stable—i.e., we can show that all of its characteristic roots are damped and decaying rather than being anti-damped and explosive. Now go back in history to 1950. The coefficients of the Europe model will be somewhat different, but suppose again that they gave rise to a stable system. Now let me give you this exact bit of information. In 1960, which is a simple mean of 1950 and 1970, by miraculous coincidence it proved to be the case that the coefficients of the model were in each and every case the ex-

act arithmetic mean of the 1950 and 1970 coefficients. What would you predict about the stability of the 1960 system?

If my asking the question had not alerted you to a paradox, I'm sure your first temptation would be to say that it is a stable system, being literally halfway between two stable systems. But that would not be consistent with Dr. Eckhaus' findings. You can make the paradox evaporate when you realize that the determinantal conditions for stability of a system [24, p. 436] do not define a stability region in terms of the coefficients of the system that is a convex region. Hence a point half-way between two points in the region may itself fall outside that region. This sort of thing does not arise in the case of well-behaved maximum systems.

I think I have said enough to demonstrate why perhaps the hardest part of my 1947 *Foundations of Economic Analysis* had to deal with the statics and dynamics of nonmaximum systems.

### Dynamics and Maximizing

Naturally this does not deny that there is a rich dynamics which can be related to maximizing. Thus consider the dynamic algorithm for finding the top of a mountain which consists of the "gradient method": this says to make your velocity in the direction of any coordinate proportional to the slope of the mountain in its direction. Such a method cannot be counted on to get you to the highest point in the Alps from any initial spot in Europe. But it is bound to converge to the maximum point of any concave surfaces that appear in the Santa Claus examples of the class-room textbooks.

Like the light rays of physics that I mentioned earlier, the optimal growth paths of the theories that have grown out of Frank Ramsey's pioneering work [19] of more than forty years ago, themselves provide a rich dynamics. Such a dynamics is quite different from that of say a positivis-

tic accelerator-multiplier analysis. You may recall that Sir William Hamilton spent a great many years trying to generalize to more than two dimensions the notion of a complex number. The story is told that his family sympathized with his earnest quest for the quaternion, and each night his children would greet him on his return from the astronomical observatory with the question: "Poppa, can you multiply your quaternions?"—only to be sadly told, "I can make my quaternions add but I can't make them multiply." Back in the 1930's if Lloyd Metzler and I had had any children, they would have asked each night: "Did all your characteristic roots turn out nicely stable?" For in those days, impressed by the stubbornness of the American Depression and its resistance to transitory pump-priming, we more or less embraced the dogma of stability.

How different were my preoccupations during the 1950's when I was on the fruitless search for a proof of the so-called "Turnpike Theorem" [26], [40], [34], [36], [37], [3]. Here one does deal with a maximizing model, at least in the sense of inter-temporal efficiency. When you study a von Neumann input-output model, it becomes the case of a min-max, or saddle-point problem like that of von Neumann's theory of games; and this destroys the possibility that your dynamic characteristic roots could all be damped. So, if my children did not treat my scholarly work with what can only be called "benign neglect," in the 1950's they would have had to ask me: "Daddy, did your characteristic roots come in reciprocal or opposite-signed pairs, as befits a catenary motion around a saddle point turnpike?"

May I crave your indulgence to digress and tell an anecdote? I do so with some trepidation because when I was invited to give this lecture I was warned by Professor Lundberg that it must be a serious one. Although it is said I was a brash young man,

I had only one encounter with the formidable John von Neumann, who of course was a giant of modern mathematics and who in addition proved himself to be a genius in his work on the hydrogen bomb, game theory, and the foundations of quantum mechanics. To illustrate his stature, I will defy Professor Lundberg even more shamelessly and tell an anecdote within an anecdote. Someone once asked Yale's great mathematician, Kakutani: "Are you a great mathematician?" Kakutani modestly replied, "Oh, not at all. I am a nothing, a mediocre plodder after truth." "Well if you're not a great mathematician, who would you name as one?" he was asked. Kakutani thought and he thought and he thought, and then according to the story he finally said—"Johnny von Neumann."

This sets the stage for my encounter with Goliath. Sometime around 1945 von Neumann gave a lecture at Harvard on his model of general equilibrium. He asserted that it involved new kinds of mathematics which had no relation to the conventional mathematics of physics and maximization. I piped up from the back of the room that I thought it was not all that different from the concept we have in economics of the opportunity-cost-frontier, in which for specified amounts of all inputs and all but one output society seeks the maximum of the remaining output. Von Neumann replied at that lightning speed which was characteristic of him: "Would you bet a cigar on that?" I am ashamed to report that for once little David retired from the field with his tail between his legs. And yet some day when I pass through Saint Peter's Gates I do think I have half a cigar still coming to me—only half because von Neumann also had a valid point.

A glance through modern journals and texts will show that, whereas the student of classical mechanics deals often with vibrations around an equilibrium, as in the case of a pendulum, the student of eco-

nomics deals more often with motions around a saddle point of catenary shape: i.e., just as a rope suspended between two nails will hang in the shape of a catenary, leaning toward the groundlevel, so will the economic motions hang in the shape of a catenary toward the turnpike. I might mention how the turnpike got its name. All Americans are used to the notion that in going from Boston to Los Angeles, the fastest way is to move quickly to a major highway and only at the end of your voyage depart to your local goal. So in economics: to develop a country most efficiently, under certain circumstances it should proceed rather quickly toward the configuration of maximum balanced growth, catch a ride so to speak on this fast turnpike, and then at the end of the twenty-year plan move off to its final goal. An interesting triple limit is involved: as the horizon becomes *large*, you spend an indefinitely *large* fraction of your time within a *small* distance of the turnpike. I shall not spell out this tongue-twister further.

### Finale

I have not been able in one lecture even to scratch the surface of the role of maximum principles in analytic economics. Nor have I even been able to present a representative sample of my own research interests in economics, or for that matter in the narrower area of maximization theory. Thus, one of my abiding concerns over the years has been the field of *welfare economics*. Along with my close friend, Abram Bergson of Harvard, I have tried to understand what it is that Adam Smith's "invisible hand" is supposed to be maximizing. Thus, consider the concept which we today call Pareto optimality—and which might with equal propriety be called Bergson optimality, since it was Bergson [2] who, back in 1938, read sense into what Pareto was groping to say and who related that narrow concept to the broader con-

cept of social norms and a welfare function. Just recently I was reading an article by a writer of the New Left. It was written in blank verse, which turns out to be an extremely inefficient medium for communication but which a dedicated scholar must be prepared to struggle through in the interest of science. The writer was scathing on the notion of Pareto optimality. Yet as I digested his message, it seemed to me that precisely in a society grown affluent, where dissident groups are called toward a way of life of their own, there arises an especial importance to the notion of giving people what *they* want. An Old Left writer dealing with a socialist economy on the verge of subsistence has surely less need for the concept of Pareto optimality than does the modern social observer in the United States or Sweden.

Moreover, it has been a special source of satisfaction to me that the calculus of modern welfare economics [31], [30], [38] was able to elucidate the old problem of Knut Wicksell [45] and Erik Lindahl [14], the analysis of public goods.

An American economist of two generations ago, H. J. Davenport, who was the best friend Thorstein Veblen ever had (Veblen actually lived for a time in Davenport's coal cellar) once said: "There is no reason why theoretical economics should be a monopoly of the reactionaries." All my life I have tried to take this warning to heart, and I dare call it to your favorable attention.

#### REFERENCES

1. G. B. Antonelli, *Teoria matematica della economica politica*, Pisa 1886.
2. A. Bergson (Burk), "A Reformulation of Certain Aspects of Welfare Economics," *Quart. J. Econ.*, Feb. 1938, 52, 310-34.
3. R. Dorfman, P. A. Samuelson, and R. M. Solow, *Linear Programming and Economic Analysis*, New York 1958.
4. R. S. Eckaus, "Dynamic Models of Domestic and International Trade," unpublished doctoral dissertation, M.I.T. 1954.
5. I. Fisher, *Mathematical Investigations in the Theory of Value and Prices*, New Haven 1925.
6. N. Georgescu-Roegen, "The Pure Theory of Consumer's Behavior," *Quart. J. Econ.*, Aug. 1936, 50, 545-93.
7. R. F. Harrod, *The Trade Cycle*, Oxford 1936.
8. J. R. Hicks and R. G. D. Allen, "A Reconsideration of the Theory of Value," *Economica*, Part 1, Feb. 1934, 14, 52-76; Part 2, May 1934, 14, 196-219.
9. H. Hotelling, "Edgeworth's Taxation Paradox and the Nature of Demand and Supply Functions," *J. Polit. Econ.*, Oct. 1932, 40, 577-616.
10. H. S. Houthakker, "Revealed Preference and the Utility Function," *Economica*, May 1950, 17, 159-74.
11. H. Jeffreys, *Theory of Probability*, Oxford 1939.
12. W. E. Johnson, "The Pure Theory of Utility Curves," *Econ. J.*, Dec. 1913, 23, 483-513.
13. J. M. Keynes, *A Treatise on Probability*, London 1921.
14. E. Lindahl, *Die Gerechtigkeit der Besteuerung*, Lund 1919.
15. E. Lundberg, *Studies in the Theory of Economic Expansion*, London 1937.
16. V. Pareto, *Manuale di economia politica*, Milan 1907.
17. ———, *Manuel d'economie politique*, Paris 1909.
18. T. Parsons, *Structure of Social Action*, 2d ed., Glencoe, Ill. 1949.
19. F. P. Ramsey, "A Mathematical Theory of Saving," *Econ. J.*, Dec. 1928, 38, 543-59.
20. ———, *Foundations of Mathematics and other Logical Essays*, London 1931.
21. P. A. Samuelson, "A Note on the Pure Theory of Consumer's Behavior," *Economica*, Feb. 1938, 5, 61-71; "—An Addendum," *Economica*, Aug. 1938, 5, 353-54. Also printed in [44].
22. ———, "The Empirical Implications of Utility Analysis," *Econometrica*, Oct. 1938, 6, 344-56; also printed in [44].
23. ———, "A Synthesis of the Principle of

- Acceleration and the Multiplier," *J. Polit. Econ.*, Dec. 1939, 47, 786-97; also printed in [44].
24. ———, *Foundations of Economic Analysis*, Cambridge, Mass. 1947.
  25. ———, "Consumption Theory in Terms of Revealed Preference," *Economica*, Nov. 1948, 15, 243-53; also printed in [44].
  26. ———, *Market Mechanisms and Maximization*, Part III, Rand Corporation, June 29, 1949; also printed in [44].
  27. ———, "The LeChatelier Principle in Linear Programming," Rand Corporation, Aug. 4, 1949; also printed in [44].
  28. ———, "The Problem of Integrability in Utility Theory," *Economica*, Nov. 1950, 17, 355-85; also printed in [44].
  29. ———, "Consumption Theorems in Terms of Overcompensation rather than Indifference Comparisons," *Economica*, Feb. 1953, 20, 1-9; also printed in [44].
  30. ———, "Diagrammatic Exposition of a Theory of Public Expenditure," *Rev. Econ. Statist.*, Nov. 1955, 37, 35-56; also printed in [44].
  31. ———, "The Pure Theory of Public Expenditure," *Rev. Econ. Statist.*, Nov. 1954, 36, 387-89; also printed in [44].
  32. ———, "Frank Knight's Theorem in Linear Programming," *Zeitschrift Für National-Ökonomie*, Aug. 1958, 18, 310-17; also printed in [44].
  33. ———, "An Extension of the LeChatelier Principle," *Econometrica*, Apr. 1960, 28, 368-79; also printed in [44].
  34. ———, "Efficient Paths of Capital Accumulation in Terms of the Calculus of Variations," in K. J. Arrow et al., eds., *Mathematical Methods in the Social Sciences*, 1959, Stanford 1960; also printed in [44].
  35. ———, "Structure of a Minimum Equilibrium System," in R. W. Pfouts, ed., *Essays in Economics and Econometrics: A Volume in Honor of Harold Hotelling*, Chapel Hill, N.C. 1960; see particularly appendix, 23-30; also printed in [44].
  36. ———, "The Two-Part Golden Rule Deduced as the Asymptotic Turnpike of Catenary Motions," *Western Econ. J.*, Mar. 1968, 6, 85-89.
  37. ———, "The Reciprocal Characteristic Root Property of Discrete-Time Maxima," *Western Econ. J.*, Mar. 1968, 6, 90-93.
  38. ———, "Pure Theory of Public Expenditure and Taxation," in J. Margolis and H. Guitton, eds., *Public Economics*, New York 1969.
  39. ———, "Stochastic Speculative Price," *Proc. Nat. Acad. Sci., U.S.*, Feb. 1971, 68, 335-37.
  40. ——— and R. M. Solow, "A Complete Capital Model Involving Heterogeneous Capital Goods," *Quart. J. Econ.*, Mar. 1956, 70, 537-62; also printed in [44].
  41. H. Schultz, *Theory and Measurement of Demand*, Chicago 1938.
  42. J. A. Schumpeter, "The Common Sense of Econometrics," *Econometrica*, Jan. 1933, 1, 5-12.
  43. E. Slutsky, "Sulla teoria del bilancio del consumatore," *Giornale degli Economisti*, 1915, 51, 19-23.
  44. J. E. Stiglitz, *The Collected Scientific Papers of Paul A. Samuelson*, Vols. I, II, Cambridge, Mass. 1965, 1966.
  45. K. Wicksell, *Finanztheoretische Untersuchungen*, Jena 1896.

# Housing Market Discrimination, Homeownership, and Savings Behavior

By JOHN F. KAIN AND JOHN M. QUIGLEY\*

The question of whether discrimination in the housing market forces Negro households to pay more than white households for identical bundles of residential services has been studied extensively. Still it remains a controversial subject. Those who claim that discrimination markups exist in urban housing markets rely principally on a series of empirical studies which conclude that blacks pay more than whites for comparable housing or that housing in the ghetto is more expensive than otherwise identical housing located outside the ghetto. (See B. Duncan and P. Hauser, R. Haugen and A. James Heins, Kain and Quigley (1970b), D. McEntire, Richard Muth, C. Rapkin, Rapkin and W. Grigsby, Ronald Ridker and John Henning, and M. Stengel.) Those who argue that price discrimination does not exist contend that studies which purport to find evidence of a discrimination markup fail to standardize completely for differences in the bundles of residential services consumed by black and white households (see Martin Bailey (1959, 1966), Richard Muth, and Anthony Pascal).<sup>1</sup> Evaluation of the diverse empiri-

cal studies leads us to conclude that blacks may pay between 5 and 10 percent more than whites in most urban areas for comparable housing. Our own analyses of a 1967 sample of nearly 1,200 dwelling units in St. Louis, Missouri suggests a discrimination markup in that city on the order of 7 percent.<sup>2</sup>

Differentials of this magnitude would represent a significant loss in Negro welfare. However, we contend that researchers, in their concern about estimating the magnitude of price discrimination, have overlooked a far more serious conse-

Bailey who concludes, "there is no indication that Negroes, as such, pay more for housing than do other people of similar density of occupation" (1966, p. 218.) A detailed presentation of the argument against the existence of a ghetto markup (price discrimination) is contained in Muth. Muth's own empirical research on the South Side of Chicago, indicates that "Negroes may pay housing prices that are from 2 to 5 percent greater" than whites pay, p. 239. However, he argues that these measured differences, and presumably those obtained in many other empirical studies, are due to cost differences (higher operating costs for housing in Negro neighborhoods) and do not represent a discriminatory markup.

<sup>2</sup> This study, based on a 1967 sample of 629 renter observations and 438 observations on single-family detached housing of the city of St. Louis, goes further than any previous study in attempting to "standardize" the bundles of residential services consumed by whites and blacks. Contract rent or market value (for owner-occupied, single-family homes) was regressed upon a detailed set of qualitative and quantitative attributes of the bundles of housing services. In addition to variables used in previous studies, such as the number of rooms and floor area, our regressions included as explanatory variables an elaborate set of quality evaluations for the dwelling unit, the structure, the adjacent structures, and the immediate neighborhood, as well as indexes of the quality of the neighborhood public school and of the level of criminal activity. Also included was a variable measuring the racial composition of the census tract in 1967. These models are discussed in some detail in Kain and Quigley (1970b).

\* Harvard University and the National Bureau of Economic Research. An earlier version of this paper was presented at the meetings of the American Economic Association in New York, December 28-30, 1969. We would like to acknowledge the assistance of Laura Stieg, Ana Bell, and William McNaught and the helpful suggestions of H. James Brown, Daniel R. Fredland, Eric A. Hanushek, Clifford R. Kern, Joseph J. Persky, T. Nicolas Tideman, and Randall D. Weiss. This paper is based on research funded by the National Bureau of Economic Research. The views are, however, those of the authors and should not be interpreted as reflecting the views of the NBER or any of its sponsors.

<sup>1</sup> The only empirical study known to us which fails to find evidence of higher prices in the ghetto is by

quence of housing market discrimination. In asking whether blacks pay more than whites for the *same* kind of housing, they have failed to consider adequately the way in which housing discrimination has affected the kinds of housing consumed by Negro households.

There is a great deal of qualitative evidence that nonwhites have difficulty in obtaining housing outside the ghetto (see James Hecht, Kain (1969), McEntire, and Karl and Alma Taeuber). Persistence, a thick skin, and a willingness to spend enormous amounts of time house-hunting are minimum requirements for nonwhites who wish to move into white neighborhoods. These psychic and transaction costs may be far more significant than out-of-pocket costs to Negroes considering a move out of the ghetto. Most blacks limit their search for housing to the ghetto; this limitation is more than geographic. There is less variety of housing services available inside the ghetto than outside; indeed, many bundles of housing services are unavailable in the ghetto at any price. This limited range of housing services within the ghetto almost certainly influences the pattern of Negro housing consumption. A full discussion and evaluation of the many ways in which discrimination may modify the housing-consumption behavior of Negro households is beyond the scope of this paper. However, the general principle can be illustrated by the differential propensities of Negro and white households to own and to purchase their homes.

Two statistical analyses of the probability of homeownership follow. The first is a detailed analysis of the probability of ownership and purchase for a sample of St. Louis households. It indicates there is a substantial difference in the probability of Negro and white homeownership and purchase even after accounting for most of the important differences in the socioeconomic characteristics of Negro and

white households. We hypothesize that this difference results from restrictions on the location and types of housing available to Negro households. This "supply restriction" hypothesis cannot be adequately tested for a single metropolitan area. Therefore, we present a second statistical analysis of the differences in actual and expected rates of Negro homeownership among eighteen large metropolitan areas. Finally, the paper examines the implications of the apparent limitations on Negro homeownership on their housing costs and capital accumulation.

### I. Homeownership and Purchase by St. Louis Households

To investigate Negro-white differences in homeownership, we developed models relating the probability of homeownership to the socioeconomic characteristics of a sample of 1,185 households in the St. Louis metropolitan area (401 black and 784 white households). Subsequently, we examined the decision to purchase or to rent for a subsample of 466 households which had changed residence in the preceding three years.

The analysis employs the regression of a binary dependent variable indicating tenure status (1 = own, 0 = rent) on several explanatory variables reflecting family size, family composition, employment status, household income, and race. Many previous studies have emphasized the importance of the family life cycle to household consumption patterns (see Martin David; John Lansing and L. Kish; Lansing, Kish, and James Morgan; Sherman Maisel; and Morgan). The life cycle hypothesis includes the combined influences of several household characteristics, which we represent by a series of family type/age interaction variables: 1) single persons (living alone or in groups) under forty-five years of age; 2) singles over forty-five years of age; 3) couples without children

with heads under forty-five years of age; 4) couples without children with heads over forty-five years of age; and 5) typical families (individuals or married couples with children).

Typical families were further described in terms of age of head, family size, number of school-age children, and by dummy variables for: female head of less than forty-five years of age; and female head of more than forty-five years of age. Income, years of schooling (of head), and number of years at present job (for head), were included as explanatory variables for all households. Race was indicated by a dummy variable (1 = Negro, 0 = white).

The probability of ownership equation, obtained by the method of generalized least squares is summarized by equation (1) (Table 1).<sup>3</sup> All the coefficients of equation (1) have the anticipated signs and are reasonable in magnitude, and twelve are highly significant statistically using conventional criteria. The results indicate that old couples are more likely to be homeowners than young couples, and old singles are more likely to be homeowners than their younger counterparts. None are as likely to be homeowners as male-headed families. Female-headed families are also less likely to be homeowners than male-headed families. Income and employment are positively related to homeownership. Family size is negatively related to homeownership, but only after adjusting for the different homeownership propensities of families with school-age children and with additional workers. The probability of ownership increases as the head of household gets older, and the introduction of a squared age term yields no evidence of any significant nonlinearity.

<sup>3</sup> The generalized least squares regression estimates are obtained by weighting each observation by  $[1/P(1-P)]^{1/2}$  where  $P$  is the value of the probability predicted by ordinary least squares. It can be shown that this procedure provides more efficient estimates of a linear probability function.

Of primary importance to this discussion is the coefficient of the race dummy variable. It indicates that, after accounting for differences in life cycle, income, education, and employment status, Negro households have a probability of ownership .09 less than that of whites. Thirty-two percent of Negro households in the sample owned their homes; if they were white, 41 percent would be homeowners.

There are some indications that the barriers to Negro occupancy in white neighborhoods are gradually declining. Thus, it could be argued that current ownership patterns primarily reflect historical discrimination and provide a misleading view of current conditions. To test this hypothesis, we estimated probability-of-purchase (1 = purchase, 0 = rent) equations for those sample households which changed their residence within the past three years. Equation (2) presents the results for the probability-of-purchase analysis (Table 1). The explanatory variables are identical to those included in equation (1). The coefficients of the dummy variables representing household type and age differ in magnitude for recent movers. Aside from these contrasts, the largest differences were obtained for the income and race variables.<sup>4</sup> The coefficient of the race dummy indicates that a Negro mover has a probability of purchase .12 lower than an otherwise identical white. Only 8 percent of Negro movers purchased homes; had they been white 20 percent would have been home-buyers.

Previous levels of housing discrimination may affect Negro households in at least one important way that is not reflected in equation (2). Because of past discrimination, Negro movers are less likely

<sup>4</sup> As with the ownership models, separate Negro and white equations were estimated for the probability of purchase. Except for their intercepts they were identical, and a covariance test indicated no statistically different relationship.

TABLE 1—MODELS OF HOMEOWNERSHIP AND PURCHASE  
Equations (1), (2), and (3) for entire sample

	Probability of Ownership		Probability of Purchase Given Move			
	Equation (1) Coef.	<i>t</i> -Ratio	Equation (2) Coef.	<i>t</i> -Ratio	Equation (3) Coef.	<i>t</i> -Ratio
Race (1=black, 0=white)	-0.088	-2.644	-0.124	-4.550	-0.091	-3.720
Income (thousands of dollars)	0.026	8.351	0.017	3.795	0.013	3.751
Years of education (head of household)	-0.006	-1.190	0.011	2.414	0.003	0.759
Years of current job (head of household)	0.002	1.908	0.001	0.624	0.002	1.365
Retired (1=yes, 0=no)	0.231	4.746	-0.070	-1.493	0.065	1.530
No household member employed (1=yes, 0=no)	-0.011	-0.277	-0.031	-0.868	-0.014	-0.362
More than one member employed (1=yes, 0=no)	0.171	6.130	-0.012	-0.438	0.002	0.102
<i>Household Types</i>						
Single females under 45 years (1=yes, 0=no)	-0.403	-5.596	-0.324	-3.665	-0.191	-2.075
Single females over 45 years	-0.295	-5.278	-0.051	-0.569	-0.183	-2.033
Single males under 45 years	-0.277	-2.577	-0.283	-2.864	-0.124	-1.170
Single males over 45 years	-0.108	-1.207	-0.057	-0.539	-0.196	-1.799
Married couples under 45 years	-0.213	-3.407	-0.290	-3.406	-0.095	-1.106
Married couples over 45 years	-0.004	-0.070	-0.124	-1.385	-0.111	-1.159
<i>Families</i>						
Age of head of household (Age) <sup>2</sup> of head of household	0.002	2.108	0.004	2.212	0.011	2.527
Number of persons (natural logarithm)	-0.156	-3.769	-0.138	-3.342	-0.113	-3.226
Number of school-age children	-0.013	-0.986	0.032	2.626	0.018	1.518
Family headed by female under 45 years (1=yes, 0=no)	-0.007	-0.148	-0.145	-3.116	-0.188	-3.712
Family headed by female over 45 years (1=yes, 0=no)	-0.192	-2.561	-0.241	-3.663	-0.206	-3.850
<i>Prior Tenure</i>						
Prior owner (1=yes, 0=no)					0.267	4.323
Prior renter (1=yes, 0=no)					0.037	1.315
New household (1=yes, 0=no)					-0.146	-3.945
Intercept	0.409	5.747	0.126	1.354	0.122	1.212
Degrees of freedom	1166		447		443	
R <sup>2</sup>	0.826		0.301		0.445	

than white movers to have been homeowners in the past. This is important because when homeowners change their residence they are more likely to buy than to rent and, conversely, when renters move they are more likely to move from one rental property to another.

In large part the association between past and present, or present and future tenure arises because renters and owners tend to differ in terms of income, family

size and composition, age, and other measured characteristics. Still, prior tenure itself may have an independent influence on subsequent tenure decisions. Therefore, probability-of-purchase equations were estimated with the addition of dummy variables for prior owner, prior renter, and new households. A fourth category "prior tenure unreported," is reflected in the intercept. Equation 3 in Table (1) illustrates these estimates.

Both the prior owner and new household variables have large and highly significant coefficients. Previous ownership raises the probability of purchase by .27. New households are .15 less likely to buy than are established households of the same age, income, and family characteristics.

Accounting for the effects of prior tenure reduces the coefficient of the race variables. Of course, the influence of housing market discrimination is reflected in prior tenure. In the sample of recent movers, only 2 percent of Negro households had previously been homeowners as compared to 17 percent of white households. Yet, even after controlling for the differences in prior tenure, Negro households are .09 less likely to become homeowners than white households in today's "open housing" market.<sup>5</sup>

Several studies of the demand for housing services have concluded that housing expenditures are more strongly related to permanent than to annual income. By extension it might be anticipated that the probabilities of homeownership and purchase would depend more on permanent than annual income. If this were true, all or part of the measured difference in the probabilities of ownership and purchase of white and nonwhite households in equations (1)–(3) might be attributable to unmeasured white/nonwhite differences in permanent incomes.

As a test of the permanent income hypothesis, we followed the convention suggested by several authors and replaced the annual income term in equations (1)–(3) by the mean incomes of the sample

households stratified by the race and by the years of education of the head of the household (see R. Ramanathan). Presumably, this averaging process reduces the transitory component of income and thereby provides an improved estimate of permanent income. The ordinary least squares estimates of the race coefficient for all three equations are consistently larger than those obtained for the current income models as are the *GLS* estimates for the purchase model. The *GLS* estimate of the race coefficient in the ownership equation using the estimated permanent income is more than twice as large as the race coefficient obtained using annual income, and the *GLS* coefficients in the purchase models using permanent income are also larger than those obtained from the equations including annual income (see equation (1), Table 2).

An alternative and more dubious (both statistically and theoretically) test of the permanent income hypothesis used an estimate of housing expenditures as a surrogate for permanent income.<sup>6</sup> The most obvious statistical problem arises because the estimate of housing expenditures for homeowners must be imputed from housing value using a gross rent multiplier. The use of different variables (monthly rent for renters and market value for owners) transformed by a constant divisor in a regression on owner/renter may produce a spurious correlation. The coefficients of this housing expenditure variable in the ownership and purchase equations vary between 20 and 45 times their standard errors; this increases our suspicion that the relation is to a significant extent spurious and arises by construction.<sup>7</sup>

<sup>5</sup> A similar difference in Negro-white probabilities of home purchase was obtained by Daniel Fredland for Philadelphia. Fredland's model differed in a number of respects from equation (3) in Table 1. It included a somewhat different set of explanatory variables, was for married households only, and was estimated by ordinary least squares. Even so, he obtained a coefficient of  $-.16$  for a minority dummy (nonwhite or Puerto Rican) and a coefficient of  $.27$  for the prior-owner dummy.

<sup>6</sup> This technique was suggested by the anonymous referee. We report it, in spite of strong statistical and theoretical reservations.

<sup>7</sup> The housing expenditure models reported in Table 2 use housing value  $\div 100$  as an estimate of homeowners'

TABLE 2—COEFFICIENTS AND *t*-RATIOS OF RACE VARIABLE FOR ALTERNATIVE SPECIFICATIONS OF INCOME: *OLS* AND *GLS*

	Probability of Ownership		Probability of Purchase			
	Equation (1)		without Prior Tenure		with Prior Tenure	
	<i>OLS</i>	<i>GLS</i>	<i>OLS</i>	<i>GLS</i>	<i>OLS</i>	<i>GLS</i>
Current Annual Income	-.150 (5.06)	-.088 (2.64)	-.154 (3.94)	-.124 (4.55)	-.114 (2.96)	-.091 (3.72)
Permanent Income	-.163 (5.23)	-.194 (6.33)	-.199 (3.65)	-.223 (4.73)	-.138 (2.58)	-.103 (2.69)
Housing Expenditure	-.048 (1.99)	-.035 (2.33)	-.077 (2.68)	-.048 (2.76)	-.069 (2.35)	-.029 (1.63)

On theoretical grounds, moreover, there is reason to suspect that even an adequate estimate of housing expenditures would not provide permanent income measures which are neutral between Negroes and whites or between homeowners and renters. If price discrimination exists in the housing market, only a demand elasticity of one for housing would prevent housing expenditures from being a biased estimate of the permanent incomes of black households. If housing demand is price elastic for blacks (see Muth), price discrimination would bias this measure of permanent income downward for blacks and would reduce the race coefficient when the ownership and purchase equations are

estimated. This bias is accentuated if housing market discrimination reduces Negro homeownership and if homeowners spend more for housing than renters of the same incomes for any reasons. Nevertheless, the race coefficients obtained from models using this estimate of housing expenditure as an explanatory variable are summarized in Table 2, which presents the race coefficients for all three alternative specifications. Also included in Table 2 are the coefficients of the racial dummy obtained by ordinary least squares.

In addition to the equations reported in Table 2, estimates were obtained employing several alternative specifications of the life cycle and age variables; tests for non-linearity in the education and income terms were also performed with negative results. For all these specifications, the magnitude and significance of the race coefficients for equations (1), (2), and (3) were virtually unchanged.<sup>8</sup>

<sup>8</sup> As a further test of the influence of housing market discrimination, separate Negro and white equations of the same form as equation (1) were estimated; a covariance test indicated no statistically significant difference between them ( $F=1.32$ ). In addition separate models for equation (1) were estimated for each of four household types described above: single persons; couples; female-headed families; and male-headed families. In each case, the coefficient of the race variable was highly significant and varied in magnitude between  $-.13$  and

monthly expenditure (see Muth). This gross rent multiplier, as well as the 1-to-120 rule (see John Shelton), is widely used in housing market analysis to make market value roughly commensurate with monthly rent. In addition to the results reported, we estimated equations using gross rent multipliers of 1/185 and 1/164. These ratios were derived by regressing monthly rent and value upon a detailed set of the individual characteristics of rental and owner-occupied units and thus deriving estimates of the equivalent value of the average rental unit ( $164 \times \text{rent}$ ) and the average rental fee for the characteristics of owner-occupied units ( $\text{value}/185$ ). The race coefficients were indistinguishable from those presented. Details of the specification and implications of these relationships may be found in Kain and Quigley (1970a and 1970b).

Taken together the estimates summarized in Table 2 and the alternatives mentioned strongly indicate that Negro households are substantially less likely to be homeowners or buyers than white households of similar characteristics. It does not "prove" that this is the result of discriminatory practices in urban housing markets; and there remain several competing explanations for these results. These alternative hypotheses may be grouped into three broad categories: 1) Differences in the "taste" for homeownership between whites and blacks; 2) Differences in the household asset and wealth positions of white and black families; and 3) Racial discrimination in the housing market as the result either of simple price discrimination in the owner and renter markets or of a more pervasive restriction on the supply of owner-occupied housing available to blacks. "Supply restrictions" could be supplemented or enforced by simple capital market discrimination or by an unwillingness on the part of banks and other mortgage lenders to finance home purchases by blacks outside the ghetto.

While it is most difficult to prove that the much lower probability of homeownership of black households is not due to differences in the taste for homeownership, many of the more commonly believed determinants of the tastes of housing consumers are included as independent variables. Furthermore, stratification by race for all of the three equations discloses no statistically significant differences. We thus conclude that the "differences in tastes" hypothesis is not an important explanation for the observed differences in market behavior between races.<sup>9</sup>

—16. When similar analyses were performed for the probability of home purchase, the sample sizes became uncomfortably small for some subgroups.

<sup>9</sup> Household expectations about moving frequency may be the only important excluded taste variable. (As will be discussed subsequently, mobility also affects the economics of homeownership.) However, a fairly

Differences in the asset or wealth positions of Negro and white households may account for part of the differences in white and nonwhite ownership and purchase probabilities. Unfortunately, the sample used in this research shares the deficiency of most other surveys in not including information on household assets and wealth. Therefore, no direct test of the asset hypothesis is possible using these data. However, for several reasons, we doubt that much of the white-Negro differences in ownership and purchase are the result of an unmeasured difference in wealth. All three equations include income, years-on-job, and life cycle variables, which may account for much of the white-black differences in assets. For most households, black and white, equity in owner-occupied housing is itself the largest component of net worth.<sup>10</sup> Therefore, in the probability of purchase model (equation (3)) prior tenure may account for much of the remaining differences in wealth. Down payment requirements are a major reason why assets might be expected to affect the decision to purchase a home. However, FHA and VA down payment requirements, especially for small single family homes purchased more than ten years ago, were small or nonexistent.

Housing market discrimination is the third, and to us, the most plausible hypothesis explaining the regression results in Table 1. The exact mechanism is hard to specify. Differential price "markups" in the owner and rental submarkets do not

extensive analysis of the mobility rates of the households in this sample indicates no important differences in the frequency of moves between white and Negro households after accounting for other socioeconomic factors.

<sup>10</sup> For example, recent Survey of Economic Opportunity tabulation indicates that for lower-middle income (\$5,000-\$7,000 per annum) families, housing equity alone represents 40 percent of the net worth of white households and an even larger proportion of the net worth of black households. See the Appendix for further details.

explain these differences in Negro and white purchase and homeownership probabilities.<sup>11</sup> We are forced to conclude that "supply restrictions" on Negro residential choice and on the kinds of housing available to black households may be largely responsible for the wide discrepancy between ownership rates for otherwise identical black and white households.

Further support for this position is provided by data on the average increase in the market value of Negro- and white-owned single family units in St. Louis. For this sample, the units owned by white central city residents have increased in value at a compound annual rate of 5.2 percent per year as contrasted to a 7.2 percent annual increase for the central city properties owned by Negro households. If this is interpreted as a difference in the net appreciation of ghetto and nonghetto properties, the findings of equations (1)–(3) become even more difficult to explain. Rather than a difference in the net appreciation of black and white owned properties, however, this finding appears to be still another manifestation of limitations on Negro residential choice. White households wishing to improve their housing can buy newer or larger houses in better neighborhoods. Negro homeowners are much less able to improve their housing in this way; as a result we hypothesize that black homeowners spend more for renovation and repair than white households of similar characteristics. An annual increase in suburban white-owned properties of 4.1 percent provides some evidence for these inferences.

<sup>11</sup> Price markups were estimated for owner and renter occupied structures using the St. Louis sample for three alternative specifications (see Kain and Quigley (1970b)). Of the three specifications, two indicate a smaller percentage markup in the owner market. Even if the markup were smaller for rental than for owner-occupied properties, it would require an extremely large price-elasticity-of-choice to reduce the probability of black ownership by 10 percentage points.

## II. Differences Among Metropolitan Areas

A complete test of the supply restriction hypothesis cannot be accomplished from an analysis of a single metropolitan area. A more powerful test of the effect of supply restrictions can be obtained by analyzing differences in black homeownership among cities. Metropolitan areas and their ghettos differ in terms of the characteristics of their housing stocks, and, therefore, in the extent to which a limitation on being able to reside outside the ghetto is an effective restriction on the supply of ownership-type housing available to blacks. For example, supply restrictions should be much less important in Los Angeles, where a large portion of the ghetto housing supply consists of single family units, than in Chicago, where ghetto neighborhoods are predominantly multi-family. We analyzed the difference between "expected" and actual black ownership rates in several metropolitan areas. Expected black ownership rates were computed by multiplying a matrix of white ownership rates (stratified into income and family size groups) by the income and family size distribution of black households. Table 3 presents this measure in 1960 for all eighteen metropolitan areas for which the necessary census data are published.<sup>12</sup> The difference between the actual black ownership rate and the expected black ownership rate for each SMSA is identical in principle to the difference in the probability of ownership attributed to race in equation (1) for St. Louis in 1967. (For St. Louis this more primitive technique yields  $-21.0$  in 1960 as compared to an *OLS* estimate of  $-15.0$

<sup>12</sup> These 18 SMSA's consisted of all those for which the data on black and white ownership rates by income and family size classes were published. The expected black ownership rate was obtained by applying the ownership proportions for white households by income and family size for each SMSA to the income and family size distribution of black households (see U.S. Bureau of the Census (1960a, Table B3)) and summing.

and a *GLS* estimate from equation (1) of -8.8 in 1967.)

As a test of the supply restriction hypothesis we then regressed these estimated differences upon 1) the proportion of central city dwelling units that are single family, a proxy for the proportion of the ghetto housing stock that is single family; 2) the proportion of the SMSA black population living in the central city, a measure of the extent of suburbanization of the black population; and 3) the actual

occupied housing or differences in the timing of urban development. Equation (4) presents the regression in difference form (expected black ownership rate minus actual black ownership rate), while equation (5) presents the same equation in ratio form (expected black ownership rate—actual black ownership rate). The *t*-ratios are in parentheses under the coefficients.

$$(4) \quad (O_B^* - O_B) = -0.24 + 0.82O_w \\ (2.36) \quad (4.64) \\ -0.36S_c + 0.12B_c \\ (6.49) \quad (2.03)$$

$$R^2 = .76$$

$$(5) \quad (O_B^*/O_B) = 0.89 + 1.52O_w \\ (1.52) \quad (1.47) \\ -1.74S_c + 0.90B_c \\ (5.34) \quad (2.52)$$

$$R^2 = .74$$

where,

$O_{B_i}^*$  = Expected black ownership rate in the *i*th SMSA

$$\left[ \sum_k \alpha_{wk_i} \cdot H_{bk_i} \right] / \sum_k H_{bk_i}$$

$O_{B_i}$  = Actual black ownership rate in the *i*th SMSA

$$\left[ \sum_k \alpha_{bk_i} \cdot H_{bk_i} \right] / \sum_k H_{bk_i}$$

$O_{w_i}$  = Actual white ownership rate in the *i*th SMSA

$$\left[ \sum_k \alpha_{wk_i} \cdot H_{wk_i} \right] / \sum_k H_{wk_i}$$

and

$\alpha_{wk_i}$  = Proportion of whites in the *k*th income/family size category who are homeowners in the *i*th SMSA

$H_{bk_i}$  = Number of black households in

TABLE 3—ACTUAL AND EXPECTED OWNERSHIP RATES OF NEGRO HOUSEHOLDS BY METROPOLITAN AREA

City	Actual	Expected
Atlanta	.31	.52
Boston	.21	.43
Chicago	.18	.47
Cleveland	.30	.58
Dallas	.39	.54
Detroit	.41	.67
Los Angeles/Long Beach	.41	.51
Newark	.24	.50
Philadelphia	.45	.66
St. Louis	.34	.55
Baltimore	.36	.61
Birmingham	.44	.56
Houston	.46	.56
Indianapolis	.45	.58
Memphis	.37	.50
New Orleans	.28	.40
Pittsburgh	.35	.59
San Francisco/Oakland	.37	.51

rate of white ownership in the SMSA.<sup>13</sup> The first two variables measure the extent of the supply restrictions among the eighteen metropolitan areas, while the latter measures any differences in the level of both white and Negro homeownership that might be attributable to such factors as intermetropolitan variation in the relative cost of owner-occupied and renter-

<sup>13</sup> The percent of single family housing in the central city for SMSA was obtained from U.S. Bureau of the Census (1960a, Table B-7). The percent of SMSA blacks residing in the central city was obtained from U.S. Bureau of the Census (1960b, Table 13).

the  $k$ th income/family size category in the  $i$ th SMSA

$S_c$  = Proportion of central city housing that is single family (Number of central city dwelling units that are single family  $\div$  total central city dwelling units)

$B_c$  = Proportion of metropolitan Negro households residing in central city (Number of Negro households in central city  $\div$  number of Negro households in SMSA).

The means and standard deviations of the variables used in equations (4) and (5) are shown in Table 4. The average ex-

TABLE 4—MEANS AND STANDARD DEVIATIONS OF VARIABLES USED IN INTER-CITY REGRESSION

	Mean	SD
$O_B^* - O_B$	0.19	0.06
$O_B^*/O_B$	1.61	0.36
$O_w$	0.65	0.07
$O_B$	0.35	0.08
$O_B^*$	0.54	0.07
$S_c$	0.55	0.22
$B_c$	0.78	0.14

pected homeownership rate for black households is .54 and the mean actual black ownership rate is .35. The actual white rate for these eighteen metropolitan areas in 1960 averages .65. Of the .30 difference between actual white and black ownership rates in these eighteen metropolitan areas Negro-white differences in family size and income account for .11; the residual difference, .19, must be attributed to other factors, including the differences in supply restrictions among the areas.

Both equations strongly support the hypothesis that the differences between observed and expected black ownership rates are small: 1) when the ghetto housing supply includes a larger proportion of single-family units; 2) when blacks have more access to the suburban housing market, with its preponderance of owner-

occupied units. As the statistics in Table 3 show, the difference between the actual and expected homeownership rate of black households is relatively small for cities like Houston and Los Angeles, where the central city and its black ghetto include more single-family housing, and is relatively large for cities like Chicago, where the ghetto is predominantly multi-family and where blacks are effectively excluded from the suburbs.

The extent of black suburbanization also appears to have a significant, though small, influence on the gap between actual and expected black homeownership. In all U.S. metropolitan areas, black households are heavily concentrated in the central cities. The mean proportion of blacks residing in the central city for the sample metropolitan areas is .78 and the standard deviation is only .14. Equation (4) indicates that a city which is one standard deviation above the mean in terms of this characteristic (92 percent of metropolitan area blacks live in the central city) would have a gap .034 larger than one which is one standard deviation below the mean (64 percent of blacks live in the central city).

The findings presented in equations (4) and (5) provide further support for the view that housing market discrimination limits Negro homeownership.<sup>14</sup> Specifically, these results indicate that a limited supply of housing suitable for homeownership in the ghetto and restrictions on Negro purchase outside the ghetto strongly affect the tenure-type of the housing consumed by Negro households as well as its location.

### III. Homeownership, Housing Costs, and Capital Accumulation

Limitations on homeownership have significant effects on Negro housing costs,

<sup>14</sup> At the minimum, it would take a peculiar spatial distribution of tastes for homeownership or of asset differences to explain these findings.

income, and welfare. As is illustrated in the Appendix, an effective limitation on homeownership can increase Negro housing costs by over 30 percent, assuming no price appreciation.

Much of the savings from homeownership results from favorable treatment accorded homeowners under the federal income tax. These tax provisions favoring homeowners are widely recognized and well documented (see Henry Aaron and Shelton). Our findings suggest that Negro households at all income levels are impeded by housing market discrimination from purchasing and owning single family homes. As a consequence, Negro households are prevented from taking full advantage of these tax benefits. Since tax savings from homeownership increase with income, this aspect of discriminatory housing markets cuts most sharply against middle and upper income black households.

Limitations on homeownership also rob Negro households of an important inflation hedge available to other low and middle income households. Calculations presented in the Appendix show that under reasonable assumptions about the appreciation of single family homes, a Negro household prevented from buying a home since 1950 would have out-of-pocket housing costs in 1970 more than twice as high as the costs which would have been incurred if the family could have purchased a home twenty years earlier.

Negro households at every income level have less wealth than white households. Current and historical limitations on homeownership may be an important reason. The importance of this method of capital accumulation among low and middle income households is apparent from a typical example. The average house purchased with an FHA 203 mortgage in 1949 had a value of \$8,286 and a mortgage of \$7,101 (see U.S. Federal Housing Administration). Assuming that this house was

purchased with a twenty-year mortgage by a thirty-year old household head, the owner of this unit would have saved more than \$7,000 and would own his home free and clear by his fiftieth birthday. Thus, if his home neither appreciated or depreciated, at age fifty he would own assets worth at least \$8,000. However, the postwar years have hardly been characterized by price neutrality. Although difficult to estimate, the average appreciation of single-family houses during the past twenty years most certainly exceeded the 100 percent increase in the Boeck composite cost index for small residential structures (see U.S. Federal Housing Administration).<sup>15</sup> This conservative 100 percent increase in value would mean that the typical FHA-financed homeowner by age fifty would have accumulated assets worth at least \$16,000, a considerable sum that he could use to reduce his housing costs, to borrow against for the college education of his children, or simply to hold for his retirement. Perspective on this hypothetical example is obtained when it is recognized that the mean wealth accumulation of white households in 1966 was only \$20,000 (see Henry Terrell). Of course, the situation would have been different if the postwar period had been one of a general decline in the price of urban real estate. But it was not.

Homeownership is clearly the most important method of wealth accumulation used by low- and middle-income families in the postwar period. Equities in single-family, owner-occupied structures account for nearly one-half of all the wealth of the lowest income group. As family income increases, the relative importance of home equities decreases. Still, home equities accounted for more than one-third of the wealth of all U.S. households earning be-

<sup>15</sup> Our sample suggests an annual rate of increase in value of white-owned properties of 4.7 percent during the 5-10 year period prior to 1966.

tween \$10–15,000 in 1962 (see D. S. Proctor et al.).

The dominant position of home equities in the asset portfolios of low and middle income households is not difficult to understand. Other forms of investment, such as the stock market, require far more knowledge, sophistication and discipline. In addition, low- and middle-income households have more leverage available in the real estate than in other investment markets.

Much of the savings imbedded in home ownership, especially among low- and middle-income households, is more or less involuntary or at least unconscious. Discipline is maintained by linking the investment (saving) decision to monthly payments for the provision of a necessity, with heavy penalties (foreclosure) imposed for failure to invest regularly.<sup>16</sup> Moreover, because of federal mortgage insurance and special advantages provided to thrift institutions, the low- and middle-income home buyer is able to borrow 90 percent or more of the purchase price of a new home. This may amount to \$15,000 or more of capital at moderate interest rates. By comparison, in the stock market he can borrow 30 percent, a ratio which he must maintain even with price declines.

If, as our findings suggest, discrimina-

<sup>16</sup> As long ago as 1953, James Duesenberry argued persuasively that levels of savings and asset accumulation are heavily dependent upon the form in which savings is maintained. Citing specifically the high proportion of savings invested in assets associated with the *reason* for saving (e.g., housing equity, pension and insurance reserves, and investment in unincorporated businesses), he suggests a close connection between the motives for saving and the form which the saving takes. Thus, although we cannot *deduce* that because people invested in some particular asset, they would not have saved if that type of asset had not been available, there appears to be a strong association.

If Duesenberry's insight is valid, then *even if* capital markets were perfect in every sense of the word, we would expect to find substantially fewer assets for households denied certain forms of saving (i.e., those forms associated with the reason for saving) such as home ownership, pension and insurance investment, and unincorporated business investment.

tion in urban housing markets has reduced Negro opportunities for homeownership, this limitation is an important explanation of the smaller quantity of assets owned by Negro households at each income level.

#### APPENDIX

##### *Owning and Saving vs. Renting and Consuming*

In an analysis of the relative costs of owning and renting a home, Shelton concludes that owning is usually cheaper than renting, as long as the household expects to live at the same location for more than three and one half years. The three and one-half year cut-off is obtained by dividing a 2 percent per year annual savings into a nonrecurring transfer cost for owner-occupied units of about 7 percent of their value.

The nonrecurring transfer cost consists of realtor commissions plus an allowance for certain fixed costs. The annual savings from homeownership include tax differences, management costs, vacancy allowances, and savings in annual maintenance expenditures for homeowners (who are able to maintain the same level of quality for about one-half percent of market value less per year). For homeowners, the total annual housing costs include: maintenance, obsolescence, property taxes, interest on mortgage, opportunity cost of money plus (discounted) transfer cost. For renters, annual rent equals landlord costs plus return on investment: maintenance, obsolescence, property taxes, vacancy allowance, management, interest on mortgage, plus return on investment.

Much of the savings from homeownership results from favorable tax provisions, i.e., from the ability to deduct interest payments and property taxes and especially from the absence of any tax on imputed rent. Therefore, the magnitude of the savings in monthly housing costs varies somewhat according to family circumstances, the size of the mortgage and the amount amortized, and the assumed opportunity costs of the family's equity.

Shelton develops an example which suggests the magnitude of the yearly savings in

housing costs obtained through ownership. This example assumes that a family may choose to buy its dwelling for \$20,000 or to rent it for \$167 per month. (This represents a gross rent of \$2,000 per year, based on a widely used gross rent/value ratio). To purchase the unit, the prospective homeowner invests \$4,000 as a down payment on the house and assumes a 6 percent mortgage.

As compared to the \$2,000 yearly rental costs, Shelton estimates that purchase would mean yearly expenses before taxes of \$1,590. Property tax and interest payments create tax shields that reduce the true costs of these two items by an amount which depends on the homeowner's tax bracket. He concludes that a conservative estimate of the tax savings created by homeownership would be \$200, yielding yearly after tax costs of ownership of \$1,390. This represents a saving of \$610 or a 15.2 percent return (after taxes) on the \$4,000 invested in homeownership, as compared to an assumed stock market return of 9 percent before taxes. Since stock market earnings are taxable, the comparable before tax return on homeownership is 18 percent. The relative return on a homeownership investment declines as the mortgage is amortized. The investment return is larger, however, if down payments are smaller or if the opportunity cost of equity capital is lower. Thus, 18 percent is likely to be a low estimate.

The savings from homeownership can also be expressed as a percentage of the costs of renting. From this viewpoint a limitation on homeownership would increase housing costs beyond three and a half years by 30 percent, assuming no price appreciation ( $\$610 \text{ savings} \div \$2,000 \text{ annual rent}$ ). As with the rate of return analysis, the savings are larger if a smaller down payment or a lower opportunity cost of capital is assumed.

Aaron obtains even larger estimates of the tax subsidy to homeowners. He presents an example, similar to the one just discussed but with a more valuable house (\$25,000) and a larger equity (\$10,000), which yields a \$342 tax saving (as contrasted to the \$200 saving computed by Shelton) and an after tax return on a \$10,000 equity of 7.4 percent

(as contrasted with a before tax return of 4 percent on other assets). However, Aaron implicitly assumes that the real price of owner- and renter-occupied housing is the same. Shelton, in contrast, contends that there is an equilibrium price difference, excluding tax differences, favoring owner-occupied housing by 1.4 percent of value. If Shelton's analysis of the comparative costs of homeownership and renting is correct in this respect, the savings to homeownership based on Aaron's example would amount to 28 percent of monthly rent computed as  $[\$342 + .014 (\$25,000 \text{ in housing value})] / (\$2,500 \text{ annual rent})$ .

The substantial divergence in housing costs noted above is in addition to any discriminatory pricing which may exist. Moreover, it must still be regarded as a lower bound estimate of the economic cost of an effective limitation on homeownership during the postwar period, since it fails to incorporate the effects of inflation on housing costs and does not admit to the special position of homeownership in the savings behavior and capital accumulation of low- and middle-income households.

A spending unit's equity in its home can be divided into three components: the initial equity or down payment, the amortization of the mortgage (savings), and any appreciation or depreciation of the property as a result of general or particular price changes (capital gains or losses). The last two items form the important link between homeownership and capital accumulation.

Although it is technically correct to view an increase in the value of an owned home as an increase in the household's wealth and to consider the opportunity cost of the equity capital as part of the spending unit's monthly housing costs, there are indications that many households do not view the matter in precisely this way. Out-of-pocket costs appear to be more important considerations for many low- and middle-income families, and it seems many view the savings in the home as a bonus to homeownership. Thus, it is of more than passing interest to compare the current out-of-pocket costs of a St. Louis family who purchased an \$8,000 FHA or VA

home on a twenty-year mortgage in 1949 with an otherwise identical family who rented throughout the entire period.

Assuming a conservative capital appreciation of 100 percent over the twenty-year period, the value of this house in 1969 would be \$16,000. Since the mortgage has been paid off, the homeowner has only insurance, real estate taxes, heating and utilities, and maintenance and repairs as out-of-pocket costs. These would total roughly \$64 per month for a St. Louis home of this value in 1969.<sup>17</sup> By comparison, a renter would have to pay somewhat more than twice this amount (\$133–160 per month) to rent a dwelling unit of this value.<sup>18</sup>

The preceding comparisons may help to explain the recent findings of the Survey of Economic Opportunity which indicate that at every level of current income, black families have fewer assets than whites, but that housing equity represents a larger proportion of the net worth of black households than of white households.<sup>19</sup> In fact, limitations on

homeownership over several generations may be an important part of the explanation of the smaller quantity of assets owned by Negro households at each income level.

#### REFERENCES

- H. Aaron, "Income Taxes and Housing," *Amer. Econ. Rev.*, Dec. 1970, 60, 789–806.  
 M. J. Bailey, "Note on the Economics of Residential Zoning and Urban Renewal," *Land Econ.*, Aug. 1959, 33, 288–92.  
 ———, "Effects of Race and Other Demographic Factors on the Values of Single-Family Homes," *Land Econ.*, May 1966, 40, 215–20.  
 G. S. Becker, *The Economics of Discrimination*, Chicago 1957.  
 M. H. David, *Family Composition and Consumption*, Amsterdam 1962.  
 J. S. Duesenberry, "The Determinants of Savings Behavior: A Summary," in W. W. Heller, F. M. Boddy, and C. L. Nelson, eds., *Savings in the Modern Economy*, Minneapolis 1953.  
 B. Duncan and P. Hauser, *Housing a Metropolis—Chicago*, Glencoe 1960.  
 D. Fredland, "Residential Mobility and

<sup>17</sup> The \$64 per month out-of-pocket costs is based on estimated homeownership costs for existing (used) FHA insured homes in St. Louis in 1967. These totaled \$58.02 for the median home in 1967 (valued at the difference in median value \$14,597 vs. \$16,000) plus increases in costs between 1967 and 1969. If the homeowner still itemizes his tax return (less likely without interest payments), he can deduct \$26 per month of these expenses. This would produce tax savings of between \$5 and \$10 per month depending on his tax bracket. These expense data were obtained from U.S. Federal Housing Administration.

<sup>18</sup> The \$133 (\$160) per month rent is again based on the same widely used rent to value ratio 1 to 120 (1 to 100). There is reason to believe the above calculations understate the extent of asset accumulation by the average homeowner. Many homeowners increase the value of their structures by improvements and additions. These outlays, of course, represent further savings and capital accumulation. Others trade up by using their accumulated equity as a down payment on a larger or better quality house, thus maintaining an even higher savings rate.

<sup>19</sup> Recent tabulations from the Survey of Economic Opportunity on the asset and liability position of Negro and white families by income group show that home equities account for an even greater share of Negro than white wealth. For example, these data indicate that white families with incomes between \$5,000 and \$7,499 have a net worth of \$12,556 as compared with a net worth of \$3,636 for Negro families in the same income class. Despite the fact that Negroes at each income level

are less likely to be homeowners, housing equity represents 67 percent of this smaller Negro net worth as compared to 40 percent of that of white families.

Although the mean housing equity of Negro homeowners is smaller than that of white homeowners, \$7,344 vs. \$11,753, the difference in Negro net worth is not to any significant degree attributable to this difference. Rather, it results from the fact that at each income level a smaller proportion of Negroes than whites are homeowners and even more importantly from the fact that the discrepancy in Negro and white ownership of other assets is even larger than the discrepancy in homeownership. Thus, if the Survey of Economic Opportunity data on assets are to be believed, Negroes in the income class \$5,000–\$7,499 have net worth in nonhousing assets equal to only 16 percent of that of white households in the same income level, and all Negroes have net worth in nonhousing assets equal to only 9 percent of that of all whites. Of course, these results can be considered as suggestive only. The weaknesses of savings and wealth data are notorious, and the interpretation of these differences, if real, would require a complete theory of Negro and white savings behavior, which encompasses the manner in which discrimination or lack of opportunity in the various markets affects the savings behavior of Negro households. The authors wish to thank Andrew Brimmer and Henry S. Terrell for making these unpublished tabulations available to them.

- Choice of Tenure," unpublished doctoral dissertation, Harvard Univ. 1970.
- R. A. Haugen and A. J. Heins, "A Market Separation Theory of Rent Differentials in Metropolitan Areas," *Quart. J. Econ.*, Nov. 1969, 83, 660-72.
- J. L. Hecht, *Because It's Right: Integration in Housing*, Boston 1970.
- J. F. Kain, "The Commuting and Residential Decisions of Central Business District Workers," in *Transportation Economics*, Universities-Nat. Bur. Econ. Res. conference series, New York 1965, pp. 245-74.
- , "Effect of Housing Market Segregation on Urban Development," *Savings and Residential Financing*, 1969 Conference Proceedings, Chicago 1969, pp. 88-113.
- and J. M. Quigley, (1970a) "Evaluating the Quality of the Residential Environment," *Environment & Planning*, Jan. 1970, 2, 23-32.
- and ———, (1970b) "Measuring the Value of Housing Quality," *J. Amer. Statist. Ass.*, June 1970, 65, 532-48.
- J. B. Lansing and L. Kish, "Family Life Cycle as an Independent Variable," *Amer. Soc. Rev.*, 1957, 22, 512-19.
- , ———, and J. N. Morgan, "Consumer Finances over the Life Cycle," in L. Clark, ed., *Consumer Behavior: The Life Cycle and Consumer Behavior*, 2, New York 1955, pp. 36-52.
- D. McEntire, *Residence and Race*, Berkeley 1960.
- S. J. Maisel, "Rates of Ownership, Mobility, and Purchase," in *Essays in Urban Land Economics*, Los Angeles 1966, pp. 76-108.
- J. N. Morgan, "Factors Related to Consumer Savings When it is Defined as a Net-Worth Concept," in L. R. Klein, ed., *Contributions of Survey Methods to Economics*, New York 1954.
- R. F. Muth, *Cities and Housing: The Spatial Pattern of Urban Residential Land Use*, Chicago 1969.
- A. H. Pascal, *The Analysis of Residential Segregation*, The RAND Corporation, P-4234, Oct. 1969.
- D. S. Projector et al, "Survey of Changes in Family Finances," Federal Reserve Tech. Paper, Washington 1968.
- C. Rapkin, "Price Discrimination Against Negroes in the Rental Housing Market," in *Essays in Urban Land Economics*, Los Angeles 1966, pp. 333-45.
- and W. Grigsby, *The Demand for Housing in Racially Mixed Areas*, Berkeley 1960.
- R. Ramanathan, "Measuring the Permanent Income of a Household: An Experiment in Methodology," *J. Polit. Econ.*, Jan. 1971, 79, 177-85.
- M. G. Reid, *Housing and Income*, Chicago 1962.
- R. G. Ridker and J. A. Henning, "The Determinants of Residential Property Values with Special Reference to Air Pollution," *Rev. Econ. Statist.*, May 1967, 49, 246-57.
- J. P. Shelton, "The Cost of Renting Versus Owning a Home," *Land Econ.*, Feb. 1968, 42, 59-72.
- M. Stengel, "Price Discrimination in the Urban Rental Housing Market," unpublished doctoral dissertation, Harvard Univ., May 1970.
- K. and A. Taeuber, *Negroes in Cities*, Chicago 1965.
- H. S. Terrell, "Wealth Accumulation of Black and White Families: The Empirical Evidence," paper presented before a joint session of the Amer. Econ. Assoc. and the Amer. Fin. Assoc., Detroit, Michigan, Dec. 28, 1970, mimeo.
- Survey Research Center, Institute for Social Research, Univ. Michigan, *1960 Survey of Consumer Finances*, Ann Arbor 1961.
- U.S. Bureau of the Census, *Statistical Abstract of the United States: 1967*, 88th ed., Washington 1967.
- , *U.S. Census of Housing: 1960, vol. II, Metropolitan Housing*, Washington 1963.
- , *U.S. Census of Population: 1960, vol. I, Characteristics of Population*, Washington 1963.
- U.S. Federal Housing Administration, *FHA Homes, 1967; Data for States and Selected Areas on Characteristics of FHA Operations under Section 203*, Washington 1967.

# Theory of the Firm Facing Uncertain Demand

By HAYNE E. LELAND\*

Traditional to economic theory is the presumption that agents act as if the environment were nonstochastic. However, recent studies of the firm facing random demand indicate that several widely accepted "truths" of the traditional theory must be abandoned. These studies nonetheless have been restrictive in two ways. Random demand has been assigned a particular functional form, most often additive or multiplicative in the random variable.<sup>1</sup> Further, these papers have not examined the differential impact of behavioral mode—the choice of price setting or quantity setting.<sup>2</sup> Although under certainty the choice of behavioral mode by a monopolistic firm is unimportant, we show that it critically conditions performance under uncertainty.

In this paper, a considerably more gen-

eral formulation of random demand is developed. This provides an analytical framework capable of examining different behavioral modes. Conditions necessary and sufficient to determine the impact of uncertainty of firms' decisions are generated. The conclusions of Baron (1970) and Sandmo for the competitive firm are shown to be special cases of more general results. The model is static, and assumes that firms produce a single output, know their cost functions with certainty, and maximize expected utility of profit.

## I. Random Demand Curves

It is necessary to formalize a notion of uncertain demand. Under certainty, the most general form of a demand relationship is an implicit function

$$(1) \quad f(p, q) = 0$$

Assuming a strictly downward sloping relationship, either  $p$  or  $q$  may be expressed as a function of the other. A natural way to introduce uncertainty is to assume the implicit demand relationship itself is random:

$$(2) \quad f(p, q, u) = 0,$$

where  $u$  is not known *ex ante*, but has subjective probability density  $dF(u)$ .<sup>3</sup> Restrictions placed on (2) are that for any  $u$ , the relation between  $p$  and  $q$  is down-

\* Assistant professor, Stanford University. This work was supported by the National Science Foundation Grant GS-3269 and GS-2530 at the Institute for Mathematical Studies in the Social Sciences at Stanford University. I wish to thank Agnar Sandmo and Bridger Mitchell for interesting discussions on the subject. An earlier version of the paper appeared in 1970, and was first presented to the Berkeley-Stanford Mathematical Economics Seminar, November 1969.

<sup>1</sup> Aspects of decision making with uncertain demand have been treated by David Baron (1970), Phoebus Dhrymes, Jacques Dreze and J. Gabsewicz, A. L. Hempenius, Ira Horowitz, Ronald McKinnon, Edwin Mills, M. Rothschild, Agnar Sandmo, Kenneth Smith, C. Tisdell, and Edward Zabel. All these have analyzed a single behavioral form; price setting, quantity setting, or price-quantity setting, and all except Horowitz assume uncertainty enters in a simple way—through a multiplicative or additive shift in the demand curve.

<sup>2</sup> A paper by Baron (1971) has just come to my attention, in which price-setting and quantity-setting behavior are considered in the same framework. Baron, however, starts with the (curious) notion that price and quantity are jointly random (see fn. 6).

<sup>3</sup> The most general possible framework would be to consider mappings from an arbitrary probability space into the set of (downward sloping) functions. Our analysis is perfectly general if the cardinality of the underlying probability space does not exceed the power of the continuum. The author thanks Karl Vind for discussions on this point.

ward sloping, and that larger values of  $u$  are associated with greater demand.<sup>4</sup> The function  $f$  is assumed to have continuous partial derivatives. These restrictions enable us to express (2) as either

$$(3) \quad p = p(q, u) \\ [\partial p(q, u)/\partial q < 0; \partial p(q, u)/\partial u > 0];$$

or<sup>5</sup>

$$(4) \quad q = q(p, u) \\ [\partial q(p, u)/\partial p < 0; \partial q(p, u)/\partial u > 0]$$

Given the cumulative function  $F(u)$ , one may derive conditional distributions  $H(q; p)$  and  $G(p; q)$ .<sup>6</sup> These distributions are symmetric:  $H(q; p) = G(p; q)$ .

A final property of stochastic demand curves proves essential to the analysis below. As total expected revenue increases (for changes in  $p$  or in  $q$ ), it seems natural to expect that the "riskiness" or dispersion of total revenue will increase. (Increased riskiness, related to stochastic dominance, is defined in the Appendix.) This *principle of increasing uncertainty (PIU)*, although perhaps not satisfied in all instances, has strong intuitive appeal. It *always* holds in the perfectly competitive case. The Appendix shows that the *PIU* is equivalent to the condition that, for all  $u$ , the sign of the partial derivative of marginal revenue with respect to  $u$  is the same

<sup>4</sup> Although this assumption is always satisfied by additive or multiplicative forms of random demand curves, it is not a weak assumption when required to hold for all  $p$  or  $q$ . It implies that we can unambiguously distinguish the favorability of a demand curve in one state of nature compared to another state of nature. Note, however, that much of the following analysis requires this condition to hold only at a single  $p$  or  $q$ , which can always be obtained by an appropriate scaling of  $u$ .

<sup>5</sup> Under perfect competition, relation (3) reduces to  $p = p(u)$ , and relation (4) does not exist.

<sup>6</sup> See Leland (1970) for a discussion of how these distributions are derived from  $F(u)$ . It is important to note that while we can derive two conditional distributions, no joint probability distribution of  $p$  and  $q$  exists. The demand *relationship* is random. A firm chooses either  $p$  or  $q$  (with certainty); the (conditional) distribution of the other variable is then uniquely determined.

as the sign of expected marginal revenue, where marginal revenue is  $\partial [p(q, u)q]/\partial q$  when quantity is fixed, and  $\partial [pq(p, u)]/\partial p$  when price is fixed.

## II. Behavioral Modes of the Firm with Stochastic Demand

When demand is random, at least four modes of behavior may be considered, depending on the flexibility of the firm to adjust output and/or price. Although these are the only control variables normally considered in the certainty situation, the firm's profit actually depends on four quantities: the price  $p$ , the demand  $q$ , the output  $X$ , and the output sold  $q_s$ . These variables must satisfy the following constraints:

$$(5) \quad f(p, q, u) = 0$$

$$(6) \quad q_s \leq q$$

$$(7) \quad q_s \leq X$$

Profit is given by

$$(8) \quad \Pi = pq_s - C(X) - F$$

where  $C(X)$  is variable cost, and  $F$  represents fixed costs.

An important consideration is whether control decisions are made before or after the resolution of  $u$ , and therefore, before or after the knowledge of the actual demand curve. Let us term *ex ante* controls those decisions made before  $u$  is known; decisions made after are termed *ex post* controls. The four modes of behavior we shall consider may be summarized in the following manner:

Description of Behavioral Mode	<i>Ex Ante</i> Controls	<i>Ex Post</i> Controls
A Certainty	—	$p, X, q_s$
B Quantity Setting	$X$	$p, q_s$
C Price Setting	$p$	$X, q_s$
D Price/Quantity Setting	$p, X$	$q_s$

It should be noted that all these controls are not independent; rather, they must satisfy constraints (5), (6), and (7).

Expected utility of profit under each behavioral mode is given respectively by

$$(9A) \quad E\left\{\max_{p, X, q_s} U[\Pi(p, q_s, X)]\right\}$$

subject to (5)–(7);

$$(9B) \quad \max_X E\left\{\max_{p, q_s} U[\Pi(p, q_s, X)]\right\}$$

subject to (5)–(7);

$$(9C) \quad \max_p E\left\{\max_{X, q_s} U[\Pi(p, q_s, X)]\right\}$$

subject to (5)–(7);

$$(9D) \quad \max_{p, X} E\left\{\max_{q_s} U[\Pi(p, q_s, X)]\right\}$$

subject to (5)–(7),

where of course the random element  $u$  is introduced into the maximization problems through the constraints. Note that

$$(10) \quad (9A) \geq [(9B), (9C)] \geq (9D)$$

That is, if a firm can choose its behavior, it will prefer total to partial to nonflexibility. What is not known a priori is the relation between (9B) and (9C): The more advantageous type of behavior appears to depend on the actual specification of random demand, cost functions, etc.<sup>7</sup>

In the certainty situation, an examination of Kuhn-Tucker conditions indicates that constraints (6) and (7) will always be met with equality. We may use these relationships to substitute for  $q_s$  and  $X$ , leaving the standard control problem in  $p$  and  $q$ , where the two are related by the demand relation (5). Because  $p$  and  $q$  decisions both are made after  $u$  is observed in the certainty situation, we may use (5) to express either control variable in terms of the other, thus generating an unconstrained problem in either  $p$  or in  $q$ —the choice of control is immaterial.

<sup>7</sup> R. Howard has suggested to me that some firms actually face different markets according to the type of behavioral mode they select. In these cases, the relationship between expected profitability of behavioral modes would be further dependent on the different market demand curves.

The equality of constraints (6) and (7) may be associated with equilibrium for consumer and firm. If the constraints are not satisfied with equality, either quantity sold is less than demand at the chosen price, implying consumer disequilibrium, or quantity sold is less than the quantity produced, implying firm disequilibrium. Under uncertainty, it will not always be optimal for these constraints to be met with equality: disequilibrium may result. But under plausible conditions, constraints (6) and (7) will be met with equality for quantity-setting and price-setting firms. This equilibrium assumption simplifies the analysis but does not affect results.<sup>8</sup>

With the equilibrium assumption, constraints (6) and (7) may be used to eliminate the variables  $q$  and  $q_s$ . Following usual notation, however, we shall call the output variable  $q$  (instead of  $X$ ); with the understanding  $q$  is determined *ex ante* for the quantity-setting firm. Equation (5) may now be used to express the *ex post* variable in terms of the *ex ante* variable and  $u$  through relation (4) or (5). After elimination, we may rewrite the problem for the quantity-setting and for the price-setting firm as

$$(9B') \quad \max_q E\{U[\Pi(q, u)]\};$$

$$(9C') \quad \max_p E\{U[\Pi(p, u)]\},$$

where

$$(11) \quad \Pi(q, u) = p(q, u)q - C(q) - F;$$

$$(12) \quad \Pi(p, u) = pq(p, u) - C[q(p, u)] - F$$

We might note that the choice of control variable ( $p$  or  $q$ ), while of no importance under certainty, is of considerable importance under uncertainty because of the asymmetry between *ex ante* and *ex post* controls. The effect of control variable

<sup>8</sup> A full discussion of this is included in Appendix B of my 1970 paper.

choice will be explored in subsequent sections.

### III. The Quantity-Setting Firm

In this section, the firm facing random demand and setting quantity is considered. The perfectly competitive firm facing a random price independent of  $q$  is a special case of this situation. In subsequent sections we study the case where the firm sets price and faces a random quantity, and the case where the firm sets both price and quantity.<sup>9</sup>

The firm seeks to

$$(13) \quad \underset{q}{\text{Maximize}} E[U(\Pi)],$$

where

$$(14) \quad \Pi = p(q, u)q - C(q) - F$$

Substituting for  $\Pi$  from (14) and differentiating (13) with respect to the control variable  $q$  yields first- and second-order conditions

$$(15) \quad E[(\partial \Pi / \partial q) U'(\Pi)] = 0,$$

or

$$E\{[p(q, u) + q[\partial p(q, u) / \partial q] - C'(q)] U'(\Pi)\} = 0,$$

or

$$E\{[MR(q, u) - MC(q)] U'(\Pi)\} = 0;$$

$$(16) \quad E[(\partial^2 \Pi / \partial q^2) U'(\Pi) + U''(\Pi)(\partial \Pi / \partial q)^2] < 0,$$

where  $U'(\Pi) = \partial U(\Pi) / \partial \Pi$ , etc. Sufficient conditions for the second derivative to be negative everywhere are that  $\Pi$  be strictly concave in  $q$  for all  $u$  and the firm be risk

neutral or risk averse ( $U''(\Pi) \leq 0$ ), or that  $\Pi$  be linear in  $q$  and the firm be risk averse. In either of these cases, a  $q$  which satisfies (15) will yield a unique global maximum of expected utility. Henceforth we assume (16) is negative for all  $q$ .<sup>10</sup>

### IV. The Effect of Uncertainty on the Optimal Output of Quantity-Setting Firms

There are several possible notions of the certainty demand curve equivalent to the random demand curve  $f(p, q, u) = 0$ . The certainty demand curve we choose to consider for the quantity-setting firm is the *expected price* certainty demand: the demand curve which would result if the firm knew price would equal its expected value with certainty for all levels of  $q$ . That is

$$(17) \quad p = E[p(q, u)] = p[q, u^0(q)] = f(q)$$

Note that  $u^0(q)$  is a function because  $p(q, u)$  is monotonically increasing in  $u$  for all  $q$ . Given this definition of *expected price* certainty demand, the linearity of the expectation and differentiation operators may be used to show that expected marginal revenue will equal the marginal revenue derived from the certainty demand curve, for all  $q$ :

$$\begin{aligned} d[qf(q)]/dq &= f(q) + q[df(q)/dq] \\ &= E[p(q, u)] + q[\partial E[p(q, u)]/\partial q] \\ (18) \quad &= E[p(q, u) + q[\partial p(q, u)/\partial q]] \\ &= E[MR(q, u)] \end{aligned}$$

Given the principle of increasing uncertainty, we can find a function  $u^1(q)$  such that

$$(19) \quad E[MR(q, u)] = MR[q, u^1(q)]$$

To examine the effect of introducing uncertainty about demand, we first ex-

<sup>9</sup> In a fully dynamic environment, the firm may exhibit various combinations of all these behavioral modes. Clearly the time span of the "static" period considered will influence the type of behavior appropriate to an industry. Agriculture firms would seem to typify quantity-setting behavior, electrical utilities price-setting behavior, and a wide variety of firms (in the short run) typify price and quantity setting behavior.

<sup>10</sup> Conditions sufficient for the existence of optimal policies are considered at length in Leland (1972).

amine the output decision of the quantity setting firm facing a certain demand curve  $p=f(q)$ .

When profit is nonrandom, maximizing profit maximizes expected utility of profit for any nonsatiated utility function. The firm with certain demand chooses  $q=q_c$  to satisfy the first-order condition

$$(20) \quad \frac{d\Pi}{dq} = \frac{d}{dq} [q_c f(q_c) - C(q_c) - F] = 0,$$

or, using (18) and (19),

$$MR(q_c, u_c^1) - MC(q_c) = 0,$$

where  $u_c^1 = u^1(q_c)$ .

Assume now that the firm does not know the demand curve  $p=f(q)$  with certainty, but rather that  $u$  is random about  $u^1(q)$  for all  $q$ . If, under uncertainty,  $q_c$  still satisfies the first-order condition (15), the introduction of uncertainty will leave optimal output decisions unchanged. If the first-order condition (15) is negative when  $q=q_c$ , the fact that second-order condition (16) is satisfied everywhere implies optimal output under uncertainty will be smaller than under certainty. If the first-order condition is positive when  $q=q_c$ , output under uncertainty will be larger.

In the analysis which follows, the effect of uncertainty depends critically on the attitude of the firm towards risk. If attitudes towards risk remain the same for all levels of profit (or wealth), we may find risk neutrality [ $U''(\Pi)=0$ ]; risk aversion [ $U''(\Pi)<0$ ], or risk preference [ $U''(\Pi)>0$ ]. Given that firms are managed according to the wishes of their owners who are typical asset holders, we might surmise that the firm will exhibit risk averse behavior.<sup>11</sup>

<sup>11</sup> Dhrymes, McKinnon, Sandmo assume risk aversion on the part of the firm; Dreze and Gabsewicz, Hempenius, Rothschild, Smith, and Tisdell assume risk neutrality, but this is perhaps to simplify their analyses, as no justification for neutrality is given. Both the notion that investors (who are risk averse) control the firm

Risk aversion implies

$$\partial[U'(\Pi)]/\partial u = U''(\Pi)(\partial\Pi/\partial u) < 0, \text{ or}$$

$$(21) \quad U'(\Pi) < U'(\Pi^1) \quad \text{for } u > u_c^1,$$

where  $\Pi^1$  is profit when  $u=u_c^1$ . From the principle of increasing uncertainty,

$$(22) \quad MR(q_c, u) > MC(q_c) \quad \text{for } u > u_c^1$$

Combining results (21) and (22) gives

$$(23) \quad [MR(q_c, u) - MC(q_c)]U'(\Pi) < [MR(q_c, u) - MC(q_c)]U'(\Pi^1),$$

for  $u > u_c^1$ . But note when  $u < u_c^1$ , both inequalities (21) and (22) are reversed, implying (23) holds for all  $u$ , or

$$(24) \quad E\{[MR(q_c, u) - MC(q_c)]U'(\Pi)\} < U'(\Pi^1)E[MR(q_c, u) - MC(q_c)] = 0$$

The right-hand side equality holds from (20), as  $E[MR(q_c, u)] = MR(q_c, u_c^1) = MC(q_c)$ .

The second-order condition implies that the first derivative of  $E[U(\Pi)]$  is monotonically decreasing in  $q$ . Therefore, the quantity  $q_a$  which satisfies (15) must be less than  $q_c$ , as the first derivative is negative when  $q=q_c$ . *The principle of increasing uncertainty implies the risk averse firm will produce less than it would under certainty.*<sup>12</sup> As noted by Sandmo, this result always

and the notion that "security" is a management goal would seem to suggest a risk averse utility function for the firm. It is of interest to note that security, a motive which cannot be incorporated in the profit-maximizing model under certainty, is introduced into expected utility of profit through the risk aversion properties of the utility function.

<sup>12</sup> The PIU is a necessary condition for points  $(q, u)$  with  $MR(q, u) > 0$  if, for any (concave) cost function and any subjective probability function  $dF(u)$ , a risk averse firm is to produce less under uncertainty. Say the PIU were violated at a point  $(q_0, u_0)$  with  $MR(q_0, u_0) > 0$ . Continuity implies the PIU will be violated within an  $\epsilon$  neighborhood of  $u_0$ . Then, for  $dF(u)$  which assigns probability one to this neighborhood, and for a cost function such that  $q_0$  is the optimal output for the certainty-equivalent demand, the presence of uncertainty will lead to a greater output.

follows for the perfectly competitive firm: In this special case, the *PIU* is always satisfied.

Risk preference implies inequality (21) will be reversed and therefore inequality (24). It follows that optimal output of the risk preferring firm will be larger under uncertainty than under certainty. Risk neutrality implies uncertainty will not affect the firm's output.<sup>13</sup>

### V. Changes in Fixed Costs

The effect of a change in fixed costs in the theory of the firm is closely related to the effect of a change in initial wealth on risky investment in portfolio theory. As the latter theory suggests, the effect depends not on the absolute level of risk aversion, but rather on the change of risk aversion as profit or wealth increases.<sup>14</sup> That is, it depends on the sign of

$$(25) \quad d[-U''(\Pi)/U'(\Pi)]/d\Pi,$$

the change in the Pratt-Arrow measure of absolute risk aversion as profits increase. It is commonly accepted that typical behavior implies the sign of (25) is negative: Investment in a risky asset is not an inferior activity. This is the "principle of decreasing absolute risk aversion."

To determine the effect of small increase  $dF$  in fixed costs, we totally differentiate the first-order condition (15), yielding

$$(26) \quad Ddq - E\{[MR(q_a, u) - MC(q_a)]U''(\Pi)\}dF = 0,$$

where  $D = \partial[E\{(MR - MC)U'(\Pi)\}]/\partial q$  at  $q = q_a$ , which by the second-order condition (16) must be negative. Solving for  $dq/dF$  gives

$$(27) \quad dq/dF = E[(MR - MC)U''(\Pi)]/D$$

<sup>13</sup> This conclusion must be modified if the "equilibrium" assumption is dropped. In that case, there are some states of nature in which the firm will not sell all its production.

<sup>14</sup> See Kenneth Arrow's 1965 study.

As  $D$  is negative, the sign of  $dq/dF$  is the opposite of the sign of  $E[(MR - MC)U''(\Pi)]$ . In a manner analogous to Sandmo, it can be shown that the principle of increasing uncertainty implies that the sign of this expectation is the opposite of the sign of (25). It follows that if absolute risk aversion decreases with profit or wealth,  $dq/dF$  is negative. *Increased fixed costs decrease output.* If absolute risk aversion is invariant to wealth, as it will be when  $U(\Pi) = c\Pi$  (risk neutrality),  $-e^{-\pi/a}$ , or positive linear transformations thereof, then  $dq/dF = 0$ . If absolute risk aversion is increasing with wealth, then  $dq/dF > 0$ .

### VI. Response to Increased Demand

When demand is random, the notion of "increased demand" is not well defined. One description, however, does suggest itself: for every state of nature, the demand curve shifts upward by an amount  $a$ . Such an increase in demand will leave the distribution of price conditional on quantity unchanged except for a higher expected value. As we wish to distinguish changing shape from upward shifts in demand, we consider the case where  $a$  is invariant to  $q$ .

Let  $p'(q, u) = p(q, u) + a$  define the demand curve resulting from the shift. The optimal output  $q$  will satisfy the new first-order condition

$$(28) \quad E\{[MR(q, u) + a - MC(q)]U'(\Pi + qa)\} = 0$$

To find the response of  $q$  to a small change in  $a$ , totally differentiate (28) and evaluate at  $a=0$  (the initial demand) and  $q=q_a$ :

$$(29) \quad dq/da = -E[U'(\Pi) + q(MR - MC)U''(\Pi)]/D,$$

where as before  $D = \partial[E(MR - MC)U'(\Pi)]/\partial q < 0$  at  $q = q_a$ ,  $a = 0$ , from the second-order condition. The change in quantity for an

increase in demand is the sum of two effects. The first,  $-E[U'(\Pi)]/D$ , is always positive, as  $E[U'(\Pi)] > 0$ ; we might term this the "revenue-substitution effect." The second,  $-qE[(MR-MC)U''(\Pi)]/D = q \cdot dq/dF$ , is positive, zero, or negative according to whether absolute risk aversion was decreasing, constant, or increasing in wealth given the *PIU*. We may term this the "risk-income effect," which normally is positive as well.

Under certainty, there is no risk-income effect: a parallel upward shift of demand will always lead to increased output. (The term also disappears when the firm is risk neutral.) Under uncertainty, increased demand may be expected to increase output, but we cannot rule out the possibility of the Giffen-type case where increased expected price produces such an increase in risk aversion that the firm chooses a smaller output.

### VII. The Price-Setting Firm

When the firm chooses to set price with uncertain demand, we may envisage two possibilities. As in behavioral mode C of Section II, the firm may have quantity flexibility, setting price and adjusting quantity to meet the actual demand. The electric power industry represents a good example of this type of behavior.<sup>15</sup> Or, as in behavioral mode D, firms may fix both price and quantity: If demand is less than output, less is sold than is produced. If demand exceeds output, there will be un-

satisfied demand. This type of behavior is explored briefly in Section IX.

Profit for the price-setting firm with quantity flexibility is given by

$$(30) \quad \Pi = pq(p, u) - C[q(p, u)] - F$$

Maximizing expected utility of  $\Pi$  with respect to the *ex ante* control  $p$  yields first- and second-order conditions

$$(31) \quad E[U'(\Pi)(\partial\Pi/\partial p)] = 0,$$

or

$$E \left\{ U'(\Pi) \left[ q(p, u) + p \frac{\partial q(p, u)}{\partial p} - C'[q(p, u)] \frac{\partial q(p, u)}{\partial p} \right] \right\} = 0;$$

$$(32) \quad \frac{\partial}{\partial p} [E[U'(\Pi)(\partial\Pi/\partial p)]] < 0$$

We assume (32) is satisfied for all  $p$ .

To examine the effect of uncertainty on the price decision made by the firm, we must again define an appropriate certainty demand curve.

Just as an *expected price* certainty demand curve was defined for the quantity setting firm, an *expected quantity* demand curve may be defined for the price setting firm. Consider the demand curve

$$(33) \quad q = h(p) = E[q(p, u)] = q[p, u^0(p)]$$

By methods similar to those in Section IV, it is readily shown that

$$\begin{aligned} (34) \quad d[p h(p)]/dp &= q(p, u) + p[\partial q(p, u)/\partial p] \\ &= E[MR(p, u)] \end{aligned}$$

The principle of increasing uncertainty, defined with respect to price changes, implies the existence of a function  $u^1(p)$  such that

$$(35) \quad E[MR(p, u)] = MR[p, u^1(p)]$$

Under certainty, the price-setting firm chooses  $p = p_0$  such that

<sup>15</sup> A simple example indicates that important differences result from a different choice of behavioral mode. Assume a demand curve given by  $f(p, q, u) = 2 - p - q + u = 0$ , where  $u = \pm 1/3$  with probability  $1/2$  each. Assume  $C(q) = q$  and the firm's utility function is logarithmic in  $\Pi$ . Under certainty, the firm would produce  $1/2$  and sell at  $3/2$ . Under uncertainty, the quantity-setting firm will sell  $(9 - \sqrt{17})/12 < 1/2$  at expected price greater than  $3/2$ . The same firm setting price, however, would choose  $p = (21 - \sqrt{17})/12 < 3/2$ , and expect to produce more than  $1/2$ . Thus the quantity-setting firm behaves *more monopolistically*, the price-setting firm *less monopolistically*.

$$(36) \quad h(p_c) + p_c[dh(p_c)/dp] - C'[h(p_c)][dh(p_c)/dp] = 0$$

To determine the effect of uncertainty *per se* on optimal price, we evaluate the left-hand side of (31) at  $p = p_c$ . The second-order condition (32) implies the optimal price will be higher, the same, or less than  $p_c$  according to whether the first-order condition evaluated at  $p_c$  is positive, zero, or negative. As with the quantity setting firm, the attitude towards risk is crucial in determining the effect of uncertainty on output. But in contrast with the quantity-setting firm, the shape of the cost curve also is critical.

*Risk neutrality* implies  $U'(\Pi) = k$  for all  $\Pi$ ; the firm maximizes expected profit.

First consider the case where marginal cost is constant:  $C'[q(p, u)] = M$  for all  $q$ . The first-order condition may be written

$$(37) \quad kE\{q(p, u) + q[\partial q(p, u)/\partial p] - M[\partial q(p, u)/\partial p]\} = 0$$

But

$$(38) \quad E\{q(p, u) + p[\partial q(p, u)/\partial p]\} = h(p) + p[dh(p)/dp],$$

and

$$(39) \quad E\{M[\partial q(p, u)/\partial p]\} = M[dh(p)/dp]$$

Substituting from (38) and (39) into (37) gives

$$(40) \quad k\{h(p) + p[dh(p)/dp] - M[dh(p)/dp]\} = 0$$

But from (36) this is satisfied by  $p = p_c$ ; the introduction of uncertainty does not affect the price decision of the price-setting risk neutral firm with constant marginal costs.

When marginal costs are not constant, the analysis is more complicated. Jensen's inequality may be used to show that if marginal cost is rising at a nondecreasing rate, the optimal price set by risk neutral

firms will be higher under uncertainty than under certainty; the opposite holds if marginal cost is decreasing at a non-increasing rate [ $C''(q) < 0$ ,  $C'''(q) \leq 0$ ].<sup>16</sup>

*Risk aversion*: Let  $p_n$  be the optimal price for the risk-neutral firm:

$$(41) \quad E[\partial \Pi(p_n, u)/\partial p] = 0$$

The optimal price  $p_a$  charged by the risk-averse firm will be greater, equal, or less than  $p_n$  according to whether the first-order condition (31) evaluated at  $p_n$  is positive, zero, or negative.

Comparative static results on the effect of risk aversion may be found for any distribution of  $u$  if  $\partial \Pi(p, u)/\partial p$  is differentiable and monotonically increasing, decreasing, or invariant with  $u$ . Assume it is one of these three. Then we may find, for  $p = p_n$ , a value of  $u_n$  such that

$$(42) \quad E[\partial \Pi(p_n, u)/\partial p] = \partial \Pi(p_n, u_n)/\partial p$$

Assume  $\partial \Pi(p_n, u)/\partial p$  is monotonically increasing in  $u$ . For  $u > u_n$ ,

$$(43) \quad \partial \Pi(p_n, u)/\partial p > \partial \Pi(p_n, u_n)/\partial p$$

But risk aversion also implies for  $u > u_n$ ,

$$(44) \quad U'[\Pi(p_n, u_n)] > U'[\Pi(p_n, u)], \text{ or } U'(\Pi_n) > U'(\Pi)$$

Therefore, for  $u > u_n$ ,

$$(45) \quad [U'(\Pi_n) - U'(\Pi)][\partial \Pi(p_n, u)/\partial p - \partial \Pi(p_n, u_n)/\partial p] > 0$$

The inequality (45) will hold for  $u < u_n$  as well, and therefore for the expectation. Sorting out the nonstochastic terms and using (42) gives

$$(46) \quad E\{U'(\Pi)[\partial \Pi(p_n, u)/\partial p]\} < 0$$

This is simply the appropriate first-order condition evaluated at  $p = p_n$ , and we saw above that in this case the optimal price  $p_a$  will be less than  $p_n$ . It follows

<sup>16</sup> A full proof is provided in Leland (1970).

immediately that if  $\partial\Pi(p_n, u)/\partial p$  is independent of  $u$ , optimal price will be unchanged from the risk neutral case, and if  $\partial\Pi(p_n, u)/\partial p$  is decreasing in  $u$ , risk aversion will lead to a *higher price*.

The preceding analysis has indicated the critical role of the change in uncertainty of profit in response to a change in price—that is, the sign of  $\partial[\partial\Pi(p_n, u)/\partial p]/\partial u$ . Unfortunately, there seems to be no clear evidence concerning this sign, nor even evidence that  $\partial\Pi(p, u)/\partial p$  is monotonic in  $u$ . Because costs were nonrandom for the quantity-setting firm, we could invoke the principle of increasing uncertainty to derive unambiguous results on the effect of uncertainty upon output. But this principle by itself is insufficient to derive comparative results for price setting firms. For example, a higher price may lead to less expected total revenue, and therefore to less uncertainty of total revenue. The same rise in price, however, may result in an offsetting change in the uncertainty of costs, leaving ambiguous the change in profit uncertainty.

Consider the following example: a firm with constant marginal costs faces random demand  $q = a + h(p)u$ , where  $u$  is a random variable with  $E(u) = 1$ . When  $a = 0$ , this is a “multiplicative” demand curve, with, standard deviation of quantity always bearing the same ratio to expected value of quantity. It is easily verified that in this case, a risk-averse firm will set a higher, equal, or lesser price than the risk neutral firm according to whether  $a > 0$ ,  $a = 0$ ,  $a < 0$ .

When the demand curve is additively separable in  $p$  and  $u$ ,  $\partial^2[q(p, u)]/\partial p \partial u = 0$ , and

$$(47) \quad \frac{\partial}{\partial u} [\partial\Pi(p, u)/\partial p] = \frac{\partial q(p, u)}{\partial u} - C''[q(p, u)] \left[ \frac{\partial q(p, u)}{\partial p} \right] \left[ \frac{\partial q(p, u)}{\partial u} \right],$$

with  $\partial q(p, u)/\partial u > 0$  and  $\partial q(p, u)/\partial p < 0$ .

*If marginal costs are nondecreasing, (47) will always be positive, implying risk aversion will lead to a lower price.* Only if marginal costs are sufficiently decreasing will (47) be negative, implying a higher price for the risk averse firm relative to the risk neutral firm. But note even here there is a problem in comparison with certainty situation, because the risk neutral firm with concave decreasing costs will set lower price under uncertainty than under certainty. When the demand curve is non-additive, the situation is even further obscured, and no relations between shapes of cost curves, risk aversion, and the effect of uncertainty are immediately apparent.

### VIII. Changes in Fixed Costs of the Price-Setting Firm

Section V examined the effect of a change in fixed costs on the output of the quantity-setting firm. A similar analysis can be used for the price-setting firm. As we might expect from above, however, the sign of the fixed cost effect is more difficult to ascertain.

Totally differentiating the first-order condition (31) gives an equation analogous to (27):

$$(48) \quad dp/dF = E[(\partial\Pi/\partial p)U''(\Pi)]/D,$$

where  $D = E\{\partial[U'(\Pi)(\partial\Pi/\partial p)]/\partial p\}$

Again we may assume decreasing absolute risk aversion with increasing wealth. In a manner precisely similar to that in Section V, it may be shown that  $dp/dF$  has the opposite sign of  $\partial[\partial\Pi(p, u)/\partial p]/\partial u$ . That is, if uncertainty of profit increases with increasing price, an increase in fixed costs will lead to a lower price, and vice versa. As above, there seem to be no a priori grounds to fix the sign of

$$\partial[\partial\Pi(p, u)/\partial p]/\partial u$$

### IX. The Price- and Quantity-Setting Firm

A final type of behavioral mode to be

considered is when the firm must select both output quantity and price *ex ante*. This is the mode first analyzed by Mills. If actual demand  $q$  is less than output, this lesser amount will be sold. If actual demand exceeds output, only the quantity produced can be sold (and at the pre-selected price). Because supply and demand will in general not be equal, this is a "disequilibrium" model, in contrast with behavioral modes A, B, and C considered above.

As before, we let  $q = q(p, u)$  represent the random demand given  $p$ . As  $q(p, u)$  is increasing with  $u$ , we may find a  $u_x$  for any  $p$  such that  $X = q(p, u_x)$ . It is readily shown that  $q(p, u) < X < q(p, \bar{u})$  when the optimal price and output are chosen, where  $\underline{u}$  and  $\bar{u}$  are the lower and upper bounds of  $u$  (perhaps  $\pm \infty$ ), respectively. Thus  $\underline{u} < u_x < \bar{u}$ .

The selection of an optimal  $p$  and  $X$  reduces to an optimal selection of  $p$  and  $u_x$ . Profit is related to these variables as follows:

$$(49) \quad \Pi = pq_s - C[q(p, u_x)],$$

where  $q_s = \min[q(p, u); q(p, u_x)]$ . Therefore

$$(50) \quad \begin{aligned} \Pi &= pq(p, u) - C[q(p, u_x)], & u \leq u_x \\ \Pi &= pq(p, u_x) - C[q(p, u_x)] \\ &\equiv \Pi_x, & u > u_x \end{aligned}$$

Differentiating (50) with respect to the control variables  $p$  and  $u_x$  gives first-order conditions

$$(51) \quad \int_{\underline{u}}^{u_x} [U'(\Pi)(\partial \Pi / \partial p)] dF(u) + \int_{u_x}^{\bar{u}} [U'(\Pi_x)(\partial \Pi_x / \partial p)] dF(u) = 0;$$

$$(52) \quad \int_{\underline{u}}^{u_x} [U'(\Pi)(\partial \Pi / \partial u_x)] dF(u) + \int_{u_x}^{\bar{u}} [U'(\Pi_x)(\partial \Pi_x / \partial u_x)] dF(u) = 0$$

Define

$$(53) \quad G(p, u_x) = E_x[U(\Pi)]$$

where  $E_x$  indicates the expectation over  $\Pi$  for  $u \leq u_x$ , and over  $\Pi_x$  for  $u \geq u_x$ . If  $G$  is a strictly concave function, implying the Hessian matrix

$$(54) \quad \begin{bmatrix} G_{pp} & G_{pu_x} \\ G_{pu_x} & G_{u_x u_x} \end{bmatrix}$$

is negative definite, the  $p, u_x$  pair satisfying (51) and (52) will represent a unique global maximum for  $G$ .

### *The Effect of Risk Aversion*

Risk neutrality implies conditions (51) and (52) reduce to

$$(55) \quad \int_{\underline{u}}^{u_x} (\partial \Pi / \partial p) dF(u) + \int_{u_x}^{\bar{u}} (\partial \Pi_x / \partial p) dF(u) = 0;$$

$$(56) \quad \int_{\underline{u}}^{u_x} (\partial \Pi / \partial u_x) dF(u) + \int_{u_x}^{\bar{u}} (\partial \Pi_x / \partial u_x) dF(u) = 0$$

Assume  $p_n, U_{xn}$  satisfy (55) and (56). We may now derive local results on the effect of risk aversion.

For small gambles, any risk-averse utility function may be approximated by the quadratic function

$$(57) \quad U(\Pi) = k\Pi - r\Pi^2$$

Clearly  $r=0$  in the case of risk neutrality. As  $r$  increases (with  $k$  remaining constant), the Pratt-Arrow measure of risk aversion will increase.

Our technique, therefore, will be to examine  $dp$  and  $du_x$  for a small increase  $dr$  in risk aversion, starting from the initial point  $p_n, u_{xn}$ , and  $r=0$ . Totally differentiating the first-order conditions gives

$$(58) \quad \begin{bmatrix} G_{pp} & G_{pu_x} \\ G_{pu_x} & G_{u_x u_x} \end{bmatrix} \begin{bmatrix} dp \\ du_x \end{bmatrix} = \begin{bmatrix} -G_{pr} dr \\ -G_{u_x r} dr \end{bmatrix}$$

where all partials of  $G \equiv E_x[k\Pi - r\Pi^2]$  are evaluated at  $p_n, u_{xn}, r=0$  and

$$(59) \quad -G_{pr} = 2 \left[ \int_{\underline{u}}^{u_x} \Pi(\partial\Pi/\partial p) dF(u) + \int_{\underline{u}}^{\bar{u}} \Pi_x(\partial\Pi_x/\partial p) dF(u) \right];$$

$$(60) \quad -G_{u_x r} = 2 \left[ \int_{\underline{u}}^{u_x} \Pi(\partial\Pi/\partial u_x) dF(u) + \left[ \int_{\underline{u}}^{\bar{u}} \Pi_x(\partial\Pi_x/\partial u_x) dF(u) \right] \right]$$

As we are starting at risk neutrality ( $r=0$ ), we may use (55) and (56) to substitute into the above expressions to give

$$(61) \quad -G_{pr} = 2 \int_{\underline{u}}^{u_x} (\Pi - \Pi_x)(\partial\Pi/\partial p) dF(u);$$

$$(62) \quad -G_{u_x r} = 2 \int_{\underline{u}}^{u_x} (\Pi - \Pi_x)(\partial\Pi/\partial u_x) dF(u)$$

First examine (62). From (50) we have  $\partial\Pi/\partial u_x = -C'[q(p, u_x)][\partial q(p, u_x)/\partial u_x]$ ,  $u < u_x$ , which is nonstochastic and negative. Removing it from the integral and noting  $\Pi < \Pi_x$  for  $u < u_x$  gives

$$(63) \quad -G_{u_x r} = 2(\partial\Pi/\partial u_x) \int_{\underline{u}}^{u_x} (\Pi - \Pi_x) dF(u) > 0$$

Fixing the sign of (61) is more tedious, and indeed possible only when  $\partial\Pi/\partial p$  is monotonic in  $u$ . In Leland (1970) it is shown that when  $\partial[\partial\Pi/\partial p]/\partial u > 0$ ,  $-G_{pr} > 0$ .

Let us assume

- (a) The demand curve  $q = q(p, u)$  is additively separable in  $p$  and  $u$ ; and
- (b) marginal cost is nondecreasing.

These assumptions may be shown to imply  $G_{pu_x} > 0$ , and that  $(\partial\Pi/\partial p)$  is increasing in  $u$  (in which case  $-G_{pr}, -G_{u_x r} < 0$ ). As the coefficient matrix in (58) is negative definite, implying negative diagonal elements, we may rewrite (58) in the form

$$(64) \quad \begin{bmatrix} - & + \\ + & - \end{bmatrix} \begin{bmatrix} \partial p/\partial r \\ \partial u_x/\partial r \end{bmatrix} = \begin{bmatrix} + \\ + \end{bmatrix}$$

The determinant of the negative definite coefficient matrix is positive, and it is readily shown that the inverse of this matrix will have all negative elements, or

$$(65) \quad \begin{bmatrix} \partial p/\partial r \\ \partial u_x/\partial r \end{bmatrix} = \begin{bmatrix} - & - \\ - & - \end{bmatrix} \begin{bmatrix} + \\ + \end{bmatrix} = \begin{bmatrix} - \\ - \end{bmatrix}$$

Therefore a small increase in risk aversion will lead the firm to *lower both price and production*, when demand is additive in  $p$  and  $u$ , and marginal costs are nondecreasing.

## X. Conclusion

We have developed a model of an expected utility maximizing firm facing random demand. The firm was assumed to produce a single product under known cost conditions. Unique to our model is the more general nature of the random demand curve. It is not restricted to a particular functional form, and it can be used to analyze different behavioral modes.

Three factors were found essential in determining the impact of uncertainty on the decisions of the firm: a) the firm's behavioral mode, as reflected by its choice of *ex ante* controls; b) the firm's attitude toward risk, as reflected by its utility function over profit; and c) the change in profit riskiness (defined in the Appendix) resulting from a change in *ex ante* controls. Given information on each of these factors, the impact of uncertainty and of parameter changes on control decisions can be determined.

The presence of uncertainty changes many of the predictions of the theory of the firm under certainty. In particular, uncertain demand implies:

- 1) The firm's selection of prices and/or outputs will not be invariant to changes in fixed costs. The results of Sandmo are extended to imperfectly com-

petitive quantity- and price-setting firms.

2) The firm in general will not be neutral between quantity-setting or price-setting behavior. If it does have flexibility to make *ex post* decisions on price or production, it will prefer either of these types of behavior to setting both price and production *ex ante*.

The effect of *risk aversion* on the decisions of a firm also was explored. Relative to a risk-neutral (expected profit maximizing) firm, risk aversion will result in:

3) A smaller output by the quantity-setting firm when the demand curve exhibits the principle of increasing uncertainty (see the Appendix).

4) Either a higher or lower price charged by the price-setting firm, depending on the change in uncertainty of profit for a change in price. Examples leading to either a higher or lower price are easily constructed. A demand curve with additive random element and nondecreasing marginal cost will lead to a lower price when the firm sets price, but also to a lower output when the firm sets quantity.

5) A lower quantity *and* price when the firm sets both *ex ante*, when risks are small, the demand curve is additively separable in  $p$  and  $u$ , and marginal costs are nondecreasing.

All these results are distribution-free, and, with the exceptions noted, independent of the functional form through which the random variable is introduced.

The model can be extended in several ways. The dimensionality of the problem could be increased in three directions: the number of inputs and outputs, the number of time periods, and the mixtures of *ex ante* and *ex post* decisions. A fully dynamic model will be necessary if inventory decisions are to be integrated with other decisions made by the firm. Zabel considers a dynamic model with multiplicative demand and risk neutrality.

It is difficult to draw conclusions regarding welfare implications of firms operating under different behavioral modes. As the example in footnote 15 indicates, the price-setting firm under some circumstances may act in a less monopolistic way than a similar firm setting quantity. On the other hand, the price-setting firm transfers some risk to factors of production, and we need a theory of optimal risk sharing to make welfare judgments on the efficiency properties of firms facing random demand.

#### APPENDIX

Consider a total revenue function  $TR(q, u)$ , where  $q$  is output and  $u$  has probability density  $dF(u)$ . Let

$$MR(q, u) = \partial[TR(q, u)]/\partial q$$

We assume  $\partial[MR(q, u)]/\partial u$  exists for all  $q$  and  $u$ .

Now let  $q$  increase by a small amount  $\delta q$ . Expected total revenue will change by

$$(A1) \quad \mu = E[MR(q, u)]\delta q$$

Assume  $\mu > 0$ . Therefore  $E[TR(q + \delta q, u)] > E[TR(q, u)]$ . The principle of increasing uncertainty states that the distribution of  $TR(q + \delta q, u)$  therefore must be riskier. *Riskier* implies that, if we were to adjust the mean of the distribution of  $TR[q + \delta q, u]$  to equal the mean of the distribution of  $TR(q, u)$  by subtracting  $\mu$  in every state of nature, then the resulting distribution of  $[TR(q + \delta q, u) - \mu]$  would be dispreferred to the distribution of  $TR(q, u)$  by any risk-averse person. That is,

$$(A2) \quad E\{U[TR(q, u)]\} - E\{U[TR(q + \delta q, u) - \mu]\} > 0$$

for any utility function with  $U''(\cdot) < 0$ . For small  $\delta q$  (and by (A1), small  $\mu$ ), we may expand  $U[TR(q + \delta q, u) - \mu]$  in a Taylor series about  $TR(q, u)$ . Omitting terms of  $O(\delta q^2)$  gives

$$(A3) \quad \begin{aligned} &U[TR(q + \delta q, u) - \mu] \\ &= U[TR(q, u)] \\ &\quad + U'[TR(q, u)][MR(q, u)\delta q - \mu] \end{aligned}$$

Using (A1) and substituting for  $U[TR(q+\delta q, u) - \mu]$  in (A2) gives

$$(A4) \quad E\{U'[TR(q, u)][MR(q, u) - E[MR(q, u)]]\delta q\} < 0$$

or

$$E\{U'[TR(q, u)][MR(q, u) - E[MR(q, u)]]\} < 0$$

We show that a necessary and sufficient condition for this to be negative for any risk averse utility function and any distribution of  $u$  is that

$$(A5) \quad \partial[MR(q, u)]/\partial u > 0$$

Sufficiency may be shown as follows. If  $MR(q, u)$  is increasing in  $u$ , there will exist a unique  $u^0$  such that

$$(A6) \quad MR[q, u^0] = E[MR(q, u)]$$

For  $\mu > u^0$ ,

$$(A7) \quad MR(q, u) - E[MR(q, u)] > 0$$

Also for  $u > u^0$ ,  $TR(q, u) > TR(q, u^0)$ , and by risk aversion

$$(A8) \quad U'[TR(q, u^0)] - U'[TR(q, u)] > 0$$

Combining these two, we have for  $u > u^0$ ,

$$(A9) \quad \{U'[TR(q, u^0)] - U'[TR(q, u)]\} \cdot [MR(q, u) - E[MR(q, u)]] > 0$$

When  $u < u^0$ , inequalities (A7) and (A8) are reversed, implying (A9) holds for all  $u$ . Taking the expectation and noting  $U'[TR(q, u^0)]$  is nonrandom gives

$$(A10) \quad E\{U'[TR(q, u)][MR(q, u) - E[MR(q, u)]]\} < 0,$$

which proves sufficiency.

Necessity (in the sense that (A2) must hold for all  $dF(u)$ ) is proved by noting that if  $MR(q, u)$  decreases with  $u$  at any value of  $u$ , we can construct a distribution  $dF(u)$  which places so much weight on this range that (A10) is not satisfied.

If  $\mu$  is negative, the principle of increasing uncertainty implies that  $TR(q+\delta q, u)$  is less risky, as  $E[TR]$  is smaller in this case. This

requires that  $\partial[MR(q, u)]/\partial u$  be negative. Therefore, the principle of increasing uncertainty is equivalent to the requirement that  $\partial[MR(q, u)]/\partial u$  has the same sign as  $E[MR(q, u)]$ .

In exactly the same manner, we may show that the PIU implies

$$\text{sign } \partial[MR(p, u)]/\partial u = \text{sign } E[MR(p, u)]$$

for all  $u$ , when the firm is a price setter and  $MR(p, u) = \partial[pq(p, u)]/\partial p$ .

## REFERENCES

- K. J. Arrow, *Aspects of the Theory of Risk-Bearing*, Helsinki 1965.
- D. Baron, "Price Uncertainty, Utility, and Industry Equilibrium in Pure Competition," *Int. Econ. Rev.*, Oct. 1970, 11, 463-80.
- , "Demand Uncertainty in Imperfect Competition," *Int. Econ. Rev.*, June 1971, 12, 196-208.
- P. Dhrymes, "On the Theory of the Monopolistic Multiproduct Firm under Uncertainty," *Int. Econ. Rev.*, 1964, 5, 239-57.
- J. Dreze and J. Gabsewicz, "Demand Fluctuations, Capacity Utilization, and Prices," CORE dis. paper 6607, Louvain 1967.
- A. L. Hempenius, "Monopoly with Random Demand," Netherlands School of Economics, Econometric Inst. rep. 6904, Jan. 1969.
- I. Horowitz, *Decision Making and the Theory of the Firm*, New York 1970.
- H. E. Leland, "On the Existence of Optimal Policies Under Uncertainty," *J. Econ. Theor.*, Feb. 1972, 4, 35-44.
- , "Theory of the Firm Facing Uncertain Demand," Institute for Mathematical Studies in the Social Sciences, Stanford University, Tech. Rep. No. 24, Jan. 1970.
- R. McKinnon, "Futures Markets, Buffer Stocks, and Income Stability for Primary Producers," *J. Polit. Econ.*, Dec. 1967, 75, 844-61.
- E. S. Mills, "Uncertainty and Price Theory," *Quart. J. Econ.*, Feb. 1959, 73, 116-29.
- R. Nelson, "Uncertainty, Prediction, and Competitive Equilibrium," *Quart. J. Econ.*, 1961, 75, 41-62.
- W. Oi, "The Desirability of Price Instability under Perfect Competition," *Econometrica*, 1961, 29, 58-64.

- J. Pratt, "Risk Aversion in the Small and in the Large," *Econometrica*, 1964, 32, 127-36.
- M. Rothschild, "The Risk Analysis of Choices Involving Uncertainty," Part II of "Essays in Economic Theory," unpublished doctoral dissertation, M.I.T., June 1969.
- and J. Stiglitz, "Increasing Risk: A Definition," *J. Econ. Theor.*, Sept. 1970, 2, 225-43.
- A. Sandmo, "On the Theory of the Competitive Firm under Price Uncertainty," *Amer. Econ. Rev.*, Mar. 1971, 61, 65-73.
- K. Smith, "The Effect of Uncertainty on Monopoly Price, Capital Stock, and Utilization of Capital," *J. Econ. Theor.*, June 1969, 1, 48-59.
- C. Tisdell, *The Theory of Price Uncertainty, Production, and Profit*, Princeton 1968.
- J. Tobin, "Liquidity Preference as Behavior Towards Risk," *Rev. Econ. Stud.*, Feb. 1958, 25, 65-86.
- E. Zabel, "Monopoly and Uncertainty," *Rev. Econ. Stud.*, Apr. 1970, 37, 205-20.

# The Administered-Price Thesis Reconfirmed

By GARDINER C. MEANS\*

The simultaneous recession and inflation of 1970 seems incapable of explanation by the textbook theories of classical economics. The substantial drop in production and the increase in unemployment has been accompanied by a substantial fall in market-dominated prices at wholesale, as is to be classically expected in a recession, but the *BLS* index of industrial prices has risen substantially and administration-dominated prices have risen even more. Theories of pricing at marginal cost or marginal revenue cannot easily explain this recession rise in administration-dominated prices. On the other hand, the administered-price thesis which challenges classical theory provides a ready explanation. This makes the validity of this thesis of immediate importance.

Recently the National Bureau of Economic Research (*NB*) brought out a report on *The Behavior of Industrial Prices* which appears to test the administered-price thesis by using newly collected price data obtained from buyers of industrial commodities.<sup>1</sup> When these valuable new data are carefully analyzed they show clear confirmation of the thesis. However, "The Main Findings" of the report, as far as they concern cyclical behavior, state that: "... we find a predominant tendency of prices to move in response to the movement of general business." And "... we find no evidence here to suggest that price rigidity or 'administration' is a significant phenomenon" (p. 9).

\* Economic consultant.

<sup>1</sup> See George J. Stigler and James K. Kindahl (S and K).

This contradiction between data and conclusions arises partly from the use of a limited version of the administered-price thesis and partly from a questionable use of the new data.

It is the purpose of this article to show that the new data overwhelmingly support the administered-price thesis; that they give the thesis a new dimension; and that the new data do not support the report's conclusion that the prices in its sample show a tendency to move with the business cycle. In order to do this, it is first necessary to examine the administered-price thesis.

## I. The Administered-Price Thesis

Basically, the administered-price thesis holds that a large body of industrial prices do not behave in the fashion that classical theory would lead one to expect. It was first developed in 1934-35 to apply to the cyclical behavior of industrial prices. It specifically held that in business recessions administered prices showed a tendency not to fall as much as market prices while the recession fall in demand worked itself out primarily through a fall in sales, production, and employment.<sup>2</sup> Similarly, since administered prices tended not to fall as much in a recession, they tended not to rise as much in recovery while rising de-

<sup>2</sup> See Means (1935). An administered price was defined as a price set for a period of time and a series of transactions. Since the early publication, the thesis has also been employed to explain the differential behavior of market and administered prices in a general demand inflation and to explain the development of inflation in the presence of a persistent excess in unemployment. (See Means (1954) (1959).)

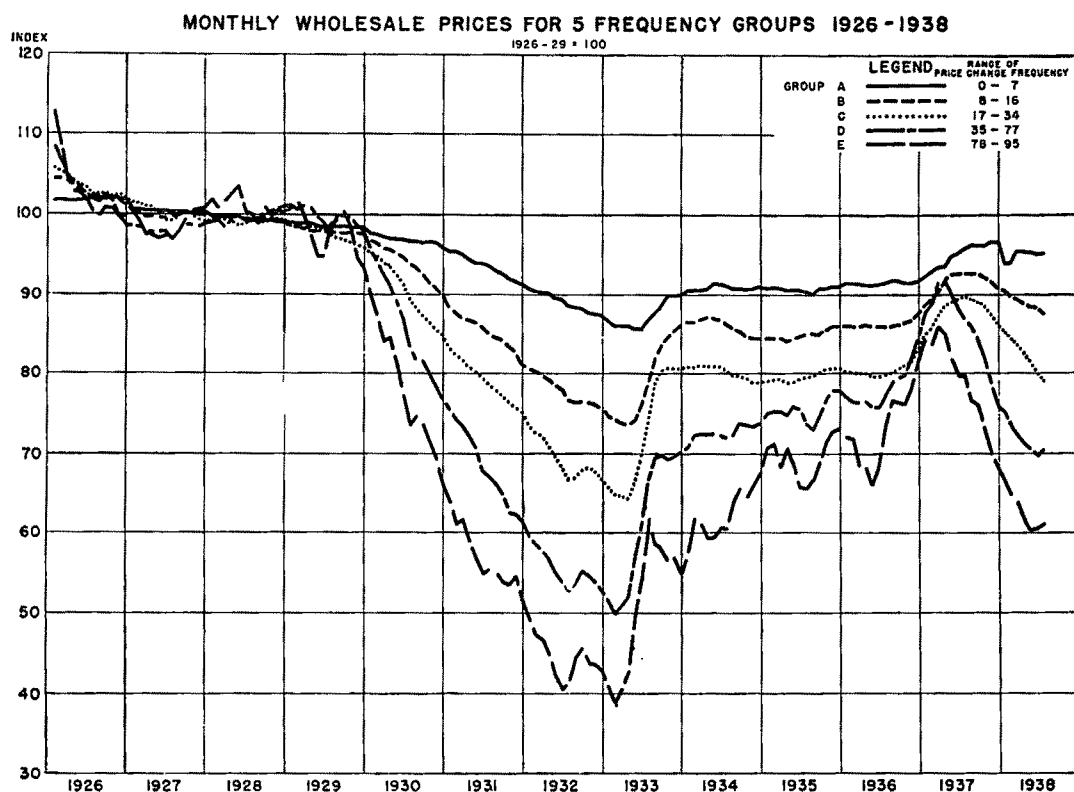


FIGURE 1

mand worked itself out primarily in a rising volume of sales, production, and employment.

This departure from classical behavior in a business cycle could theoretically take any one of three forms. In a recession an administered price might fall substantially less than classically competitive market prices; it might show no substantial change; or it might rise contracyclically. These can be referred to, respectively, as relatively inflexible, rigid, and contracyclical behavior. Any one of these three reactions to a general fall in demand would be classically unexpected except as some noncyclical factor intervened such as a trend of technical change. Similarly, in a recovery, an administered price might rise less, show no change, or actually fall.

In the early development of the admin-

istered-price thesis, the main focus was on relative inflexibility, not on price rigidity or contracyclical behavior. This is clearly indicated in the title of Senate Document 13, *Industrial Prices and Their Relative Inflexibility*. It is also clear in the crucial chart published in *The Structure of the American Economy* which clarifies the thesis and is reproduced here as Figure 1.<sup>3</sup>

<sup>3</sup> The validity of this figure and other data I have published on frequency of price change has been challenged by the authors. They say "The McAllister analysis effectively destroys the entire body of work resting on frequency of price change" (see Stigler and Kindahl, p. 20), and apply this statement to my "original work." But the McAllister analysis bears no relation to my frequency analysis as Stigler now recognizes. McAllister established the mathematically obvious fact that where there are several separate companies reporting, each changing its price infrequently, an index combining their reports would be likely to change more frequently than the average frequency of change of the separate

As can be seen it is wholly concerned with the relative inflexibility of administered prices. In neither report is attention given to rigid or contracyclical price behavior.

This preamble on the administered-price thesis is necessary because the authors of the National Bureau report have confused the issues by attributing to me a very partial version of the administered-price thesis which takes no account of those administered-prices which show *relative* inflexibility. Thus, on its first page, the report speaks of the early administered-price investigations in these terms: "The inquirer was Dr. Gardiner Means and his famous answer was that large numbers of industrial prices were *wholly* unresponsive to cyclical fluctuations of markets" (p. 3, emphasis added). No citation is given for this statement and it is contrary to both the evidence I presented and the thesis I developed.

Throughout the first two introductory chapters of the report, the idea is fostered that the administered-price thesis has nothing to do with relative inflexibility. Thus the report says "The main thrust of the doctrine of administered prices is that contractions in business lead to *no* system-

atic reduction of industrial prices, and much more equivocally, expansions in business may only tardily lead to price increases" (p. 7).<sup>4</sup> Further, the report refers to the "Means' " statistics as "These startling statistics of price rigidity—" (p. 13).

Nor can there be any question that the authors are aiming to test the Means' thesis. Means is indicated as the source of the "doctrine" being tested. The name "Means" appears seventeen times in the first eighteen pages. And no other source is given for the doctrine.

The partial character of the thesis being tested is also apparent in the classification used in the master table summarizing the cyclical results of analysis. It classifies commodity prices according to whether, in a recession, they decreased, showed no change, or increased.<sup>5</sup> It treats *all* price decreases as in conflict with the administered-price thesis even if the decline is relatively small compared with the decline in competitive market prices. Thus, it leaves out an essential part of the administered-price thesis.<sup>6</sup>

<sup>4</sup> The "equivocal" factor is introduced by the authors' failure to distinguish between business recovery and business expansion. The administered-price thesis points to one type of expectation where there is a business recovery following an immediately preceding recession and another where business expansion does not follow an immediately preceding recession.

<sup>5</sup> "No change" is defined to include all prices that change .05 percent a month or 1/2 percent or less in a ten-month recession.

<sup>6</sup> What is being tested still remains ambiguous since nowhere in the text are the figures for price rigidity and for contracyclical behavior combined in a total and discussed as they support or contradict the thesis the authors indicate they are testing. Rather, having asked, "What is the verdict of the present price data?" the authors answer with respect to the two business contractions studied: "There are two NB price movements in the expected fashion for each perverse change" (see S and K, p. 61). For both contractions and expansions they say, "Our general conclusion, then, is that the behavior of industrial prices is not perverse" (p. 63). To establish that more industrial prices go down in a contraction *than go up* is not even a test of the thesis propounded in the text. Yet it is the only specific application of the data in the master tabulation to test a thesis other than the classical.

---

companies. Both McAllister and Stigler and Kindahl assume that the frequency figures published in my early work were obtained by counting the number of times the *published price indexes* changed. Actually, the frequency counts given in my publications were based on the raw data supplied to the BLS by the separate reporting companies, and the frequency counts in *The Structure of the American Economy* show the average of the frequency of change of the separate reporters, as is there indicated (see p. 187). Stigler has recently written (1971) for the record:

The criticism that Dr. Means' basic work in the 1930's ignored the number of price reporters in calculating the frequency of price changes (*The Behavior of Industrial Prices*, p. 20) is based upon a misunderstanding of his procedures, and I apologize for the error (which I had previously committed in the *Journal of Business*, Jan. 1962, p. 5). He informs me that he used the average number of price changes per reporter in the work reported in the *Structure of the American Economy*, Part I (Washington 1939) and an approximation to this in *Industrial Prices and Their Relative Inflexibility* (Washington, 1935)... [p. 852]

Hereafter this partial thesis will be referred to as the truncated version of the administered-price thesis. Of course, if the statistics support this truncated version, and this will be shown to be the case, they also support the full administered-price thesis although the reverse does not follow.

In examining the actual data the full administered-price thesis and the truncated version must both be kept in mind and set against the classical thesis which the report states in the following words: "Classical theory leads one to expect prices to fall in competitive industries during a business contraction, because both demand and marginal costs fell, and the reverse movements will occur in expansions" (S and K, p. 60).<sup>7</sup>

In what follows, the new National Bureau price data will be employed to test the administered-price thesis, and to expand its dimensions. Then the reasons for the discrepancy between the report's conclusions and its data will be examined. Finally, the essentials of the challenge to classical theory implicit in the new data will be considered.

## II. Price Behavior in Two Cycles

As can be seen in Figure 2, the Federal Reserve Board index of industrial production for the period covered by the National Bureau report shows only two clear cycles in which there is a sharp recession followed by a sharp recovery. The first runs twenty-three months from July 1957 to June 1959 and the second runs from January 1960 to early 1962. The exact month when the sharp rise in the second cycle comes to an end is not clear since the sharp rise turned into a continuing slow rise with no downturn. Here it will be assumed that it terminated when industrial recovery had exceeded the industrial recession by 50 per-

<sup>7</sup> It should be noted that an administered price which fell much less than called for by the fall in demand and marginal costs would, by this statement, be classed as behaving classically.

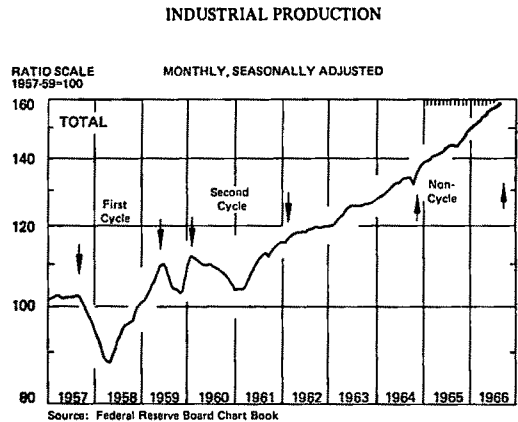


FIGURE 2

cent, the same proportion that occurred in the first cycle. This gives March 1962 as the end of the sharp upturn and provides a complete cycle of twenty-six months. The new *NB* data for these periods will be used to test the administered-price thesis, first in terms of the truncated thesis, for which the data were classified in the report, and then in terms of the full thesis.

The National Bureau report contains monthly price indexes covering the relevant periods for sixty-three commodities. Of these, thirteen should be classed as market-dominated, leaving fifty commodities which meet the requirements of the administered-price thesis and can provide a test of that thesis.<sup>8</sup>

A compilation of the behavior of the fifty commodities in each of the two cycles is shown in Table 1.<sup>9</sup> In 69 percent of the 200 opportunities to change, price behavior was consistent with the truncated administered-price thesis.

However, the real test of the truncated administered-price thesis is the behavior

<sup>8</sup> The fifty commodities are listed in Appendix A and the commodities classed here as market-dominated are listed in Appendix B. The latter list includes prices which were made in highly competitive markets such as plywood, wood flooring and copper as well as products such as copper tubing which were simply fabricated from highly competitive raw materials.

<sup>9</sup> The behavior of the individual indexes is given in Appendix A.

TABLE 1

Number of Indexes	1st	1st	2d	2d	Total Oppor- tunities for Change
	Contraction July 1957 to Apr. 1958	Recovery Apr. 1958 to June 1959	Contraction Jan. 1960 to Jan. 1961	Recovery Jan. 1961 to Mar. 1962	
Conforming to the Truncated Administered-Price Thesis	34	33	26	45	138
Not Conforming	16	17	24	5	62
Total	50	50	50	50	200
Percent Conforming	68	66	52	90	69

of each index for the combination of the two cycles. The result of a compilation of such behavior is given in Table 2 under the assumption that an index which conforms to the expectation of the truncated thesis three or four times out of the four movements tends to support that thesis. Here we have fifteen times as many administration-dominated indexes which show a tendency to support the truncated administered-price thesis as shows the opposite tendency.<sup>10</sup>

Even this does not give the complete test of the administered-price thesis since relative change is not taken into account. In the two contractions the eighteen indexes which showed no tendency either way dropped an average of 2.0 percent compared with a 6.7 percent average drop for the thirteen market-dominated indexes in the *NB* sample. In the two recoveries, the eighteen neutral indexes showed an average *drop* of .8 percent compared with a 3.5 percent average rise for the thirteen market-dominated indexes. Thus, while the new data give strong support for

the truncated thesis, they give even greater support for the full administered-price thesis. Only to a nonsignificant extent did the administration-dominated prices fail to conform to the expectation of the administered-price thesis.

### III. Behavior in Special Cycles

The most direct evidence of nonclassical behavior given in the report is buried in an appendix which examines what the authors call "Specific Cycles." The authors found sixty-six cases during their test period in which the output in an industry or product changed at least 20 percent in a period of

TABLE 2

	Number of Indexes	Percent
Conforming to the Truncated Administered-Price Thesis		
In all 4 movements	10	
In 3 of 4 movements	20	
	—	
Tending to support the Truncated Administered-Price Thesis	30	60
Not Conforming to the Truncated Administered-Price Thesis		
In all 4 movements	None	
In 3 of 4 movements	2	
	—	
Tending not to Support the Truncated Administered-Price Thesis	2	4
Neutral in Tendency —(2 each way)	18	36
Total	50	100

<sup>10</sup> The term "market-dominated price" is used here to include not only prices which are made in the market by the interaction of a large number of buyers and sellers but also administered prices in which one or more market-priced raw materials constitutes a substantial part of the product's cost and leads to frequent adjustments in price. The term administration-dominated price covers administered prices which are not appropriately classed as market-dominated because of market-priced raw materials.

eight to ten months and in which they expected sharp short-run fluctuations in output of these durations to be dominated by demand changes.<sup>11</sup> With this magnitude of change in production brought about predominantly by a change in demand, classical theory would lead one to expect a substantial price change in the same direction.

The actual results of this test are dispersed in an appendix table which compares the behavior of the *NB* and the *BLS* indexes for conformity to the classical expectation. When the *NB* findings are separated out and combined (see Table 3), they show<sup>12</sup> that in at least 85 percent of the cases, the data on specific cycles conform to the administered-price thesis.<sup>13</sup> Again this is overwhelming support for the truncated administered-price thesis and therefore overwhelming support for the full administered-price thesis as they apply to specific cycles.

TABLE 3—PRICE BEHAVIOR IN SPECIFIC CYCLES

	Prices Showing Pro-cyclical Behavior	Prices Showing No Change or Contra-cyclical Behavior	Percent Conforming to the Administered-Price Thesis
In 10 Specific Contractions	1	9	90
In 56 Specific Expansions	9	47	84
	10	56	85

Strangely, neither this test nor its remarkable result is mentioned in "The Main Findings" of the *NB* report (see pp 7-10). The test is mentioned in the

main text but as a "... relatively unsuccessful investigation. . . ." (p. 44) and relegated to a footnote on page 64 and an appendix. However, the conclusion is reached in the appendix that "A tabulation of price movements against output movements in specific cycles leads to unprepossessing results; in general, neither price [the *NB* index and the comparable *BLS* index] conforms to output changes" (p. 194). No explicit indication is given as to why these "unprepossessing results" are not made a part of "The Main Findings."

#### IV. A New Dimension

While the original administered-price thesis was only concerned with a failure to conform to classical expectations, actual experience with industrial prices in recent years has disclosed many cyclical cases in which price behavior has been the *reverse* of that to be expected from classical theory, the price rising with recession and falling with recovery.

Such contracyclical behavior is evident in the new *NB* commodity indexes. Of the fifty administered-price indexes which are relevant to the administered-price thesis, forty-six show contracyclical behavior in at least one of the four opportunities available in the two industrial recessions and two recoveries shown in Figure 2. Classing the indexes by behavior we get the following:<sup>14</sup>

Contracyclical Behavior in all 4 opportunities	2
Contracyclical Behavior in 3 of 4 opportunities	8
Contracyclical Behavior in 2 of 4 opportunities	19
Contracyclical Behavior in 1 of 4 opportunities	17
No Contracyclical Behavior	4
	50

<sup>11</sup> See S and K, Appendix D, pp. 193-96.

<sup>12</sup> See S and K, p. 194. For the purpose of the test, the authors classed a price index which changed less than 2 percent in ten months as showing "no change."

<sup>13</sup> The actual commodities involved in this test are not given and it might be that some market-dominated items are included in the ten items which conformed to the classical expectation.

<sup>14</sup> Derived from table, Appendix A.

Altogether these fifty *NB* "commodity" indexes behave in a contracyclical manner in 87 of the 200 possible movements or 43.5 percent of the opportunities. Only four of the indexes show no case of contracyclical behavior.

### V. The Stigler-Kindahl Conclusion

With this statistical data available to the authors of the report, just what is the basis for their "main finding" on "Cyclical Behavior" that "... we find a predominant tendency of prices to move in response to the movement of general business" (p. 9)?

The master tabulation of the new data with which they support this main finding is given in Tables 4 and 5.<sup>15</sup>

TABLE 4—THE DIRECTION OF PRICE CHANGE IN TWO CONTRACTIONS

Price Changes	All Prices	Excluding Steel Prices
Decreases	40	40
No Change <sup>a</sup>	10	7
Increases	18	10
Total	68	57
Proportion Decreasing (percent)	59	70

<sup>a</sup> —.05 to +.05 percent per month.

When we look at the figures for all prices in the sample we find that in only 77 cases out of 138 or 56 percent does price move in the same direction as business. In itself, this is not a very robust support for a finding of a *predominant tendency* to so move. More support is apparent when steel prices are excluded since their behavior clearly conforms to the expectation of the administered-price thesis even in its truncated form.

The authors justify the exclusion of the

TABLE 5—DIRECTION OF PRICE CHANGE IN TWO EXPANSIONS

Price Changes	All Prices	Excluding Steel Prices
Increases	37	36
No Change <sup>a</sup>	14	13
Decreases	19	10
Total	70	59
Proportion Increasing (percent)	53	61

<sup>a</sup> —.05 to +.05 percent per month.

steel data on the ground that they "... are numerous in our sample and atypical in their price behavior ..." (p. 8). But frequency in the sample alone is no basis for exclusion and no evidence is given that steel prices are atypical. Actually there is substantial evidence that their behavior is quite typical of that of other administration-dominated prices. For example, in the two cycles of Figure 2, the nine steel price indexes did not move *with* business in 74 percent of the cyclical movements while the more numerous chemicals (22) did not move with business in 75 percent of the occasions and half of the remaining nineteen administered prices show 75 percent or more nonconformity. If there is atypicality in the sample it is the behavior of the market-dominated prices which are included. Thus, the compilation excluding steel would seem to give no relevant support to the "predominant-tendency" conclusion. It must find its support in the total sample.

A second questionable factor is the use of the economic expansion from November 1964 to November 1966 as a period of *cyclical* movement. This is described in the text as a "... short, sharp expansion ..." but also as part of an expansion "... so long that it partakes of a trend ..." (p. 8). Examination of Figure 2 shows no short,

<sup>15</sup> See S and K, pp. 8-9.

sharp rise in the index of industrial production for the 1964 to 1966 period covered in the report. By early 1962 production had recovered from the *cyclical* fall of 1960 and in no sense did the 1964-66 rise follow an immediately preceding recession. Rather it followed a period of relative stagnation and both the administered-price thesis and the classical thesis point to an expectation that under such conditions industrial prices would rise with industrial activity. Thus, a finding of a predominant tendency for prices to rise in this period could not possibly support either thesis as against the other.

Because no separate figures are given for the two expansions, it is necessary to regard the combined expansion figures as irrelevant to a cyclical test since one set is noncyclical. The need for excluding the combined figures is heightened when we read that "... cyclical conformity was only fair in the 1958 to 1960 expansion but very good in the 1964 to 1966 expansion" (S and K, p. 63), a result quite in conformity with administered-price expectations. This reduces the relevant evidence in the master table to that for all prices and to their behavior in the two recessions.

The all-price data in the two recessions indicate that 59 percent of the prices fell with the decline in business but this result, favorable to the classical expectation, disappears when adjustments are made for trend. When there is a strong downward trend in price, a small fall in a recession may actually reflect a cyclical rise. Thus the report says, "When there are strong trends in the price indexes, the cyclical responsiveness of price may be swamped or exaggerated by the trend component" (p. 44). Fortunately the report contains a tabulation corrected for trend (p. 62). The authors' unadjusted and their trend-adjusted figures are given below for the recession behavior of all prices in Table 6. The trend-corrected data do not show *any*

TABLE 6—PRICE BEHAVIOR IN TWO RECESSIONS

Price Changes	Raw Data	Data Corrected for trend
Decrease	40	32
No Change <sup>a</sup>	10	24
Increases	18	12
Total	68	68
Percent Decreasing	59	47

<sup>a</sup> - .05 to +.05 percent per month.

tendency to conform to classical expectations, let alone a predominant tendency.

Both logic and the authors' own statement on trend suggest that the trend-adjusted figures, not the unadjusted, are the figures that are relevant to the master table and the main finding on cyclical behavior.

It has already been indicated that the *NB* sample contains a number of market-dominated prices. Of the sixty-three commodities for which relevant data is given, thirteen must be classed as market-dominated. (See Appendix B.) Of the five included in the compilation above for which no data is given, two appear to be market-dominated.<sup>16</sup> This means that fifteen commodities need to be removed from the sixty-eight in the above tabulation if inferences are to be drawn from it as to the cyclical behavior of administration-dominated prices.

Examination of the data given in the report indicates that, as behavior is measured in the master table, all of the thirteen market prices for which data is given not only showed decreases in the two recessions but would have shown decreases if the price series had been corrected for trend. Independent evidence also indicates that this would almost certainly be true of the two for which no data was given. This sug-

<sup>16</sup> Copper sheet and strip; Brass sheet and strip.

gests that fifteen of the items classed as "decreases" should be eliminated if the data is to be used to cast light on the behavior of administered prices. Removing them would result in the tabulation shown in Table 7.

TABLE 7—BEHAVIOR OF ADMINISTRATION-DOMINATED PRICES IN TWO RECESSIONS

Price Change	Data Uncorrected, for Trend	Data Corrected for Trend
Decreases	25	17
No Change <sup>a</sup>	10	24
Increases	18	12
Total	53	53
Percent Decreases	47	32

<sup>a</sup> -.05 to +.05 percent per month.

One further correction needs to be made in the data if individual items are to be used to measure a cyclical tendency. In the master table, the authors have classed individual items according to their *average* behavior in the two recessions. As a result, a price which rose in one recession but fell somewhat more in the other is classed as falling in the two recessions. Likewise a price which showed no change in one recession and declined in the other is also classed as declining. Certainly in neither case would the price show a *tendency* to decrease. And more important it would confirm the administered-price thesis (in its truncated form) just as much as it confirmed the classical thesis or, more correctly, it would be neutral, confirming neither.

Nowhere in the report is this averaging procedure indicated, yet a classification which separated the behavior in the two recessions would present a significantly different picture. At least eleven of the twenty-five administration-dominated prices classed as decreasing in the two recessions according to the unadjusted data and at

least ten of the seventeen in the trend-adjusted data actually went down in only one of the two recessions, either going up or showing no change in the other. Eliminating these gives a final tabulation shown in Table 8. Thus, even including the fifteen

TABLE 8—BEHAVIOR OF PRICES IN TWO RECESSIONS (Data corrected for trend)

Price Change	Total Indexes in <i>NBER</i> Sample	Administra- tion- Dominated Indexes in <i>NBER</i> Sample
Decreases in Both Recessions	22	7
Not Decreasing in Both Recessions	46	46
Total Commodities	68	53
Percent Decreasing in Both Recessions	32	13

market-dominated prices, only 22 or 32 percent of the total National Bureau sample give actual support for the authors' finding of "a predominant tendency of prices to move in response to the movement of general business." And more important for the "doctrine of administered-prices," only seven of the administration-dominated prices or 13 percent show a *tendency* to behave in a fashion different from that to be expected under the administered-price thesis in its truncated form. Presumably even a smaller percentage would show procyclical movement in three or four of the possible changes if two full cycles were covered, and of these some could be expected to show relative inflexibility.

The conclusion seems justified that the data presented by the authors in no way support their main finding that the industrial prices of their sample show a predominant tendency to behave in a classical

fashion or challenge the administered-price thesis. Rather, their statistical evidence, more carefully examined, constitutes a major challenge to classical theory.

## VI. The Essence of the Challenge to Classical Theory

That the cyclical behavior of administered prices presents a basic challenge to classical theory is clear. But the essence of that challenge must be found in the specific transaction prices received by specific sellers or paid by specific buyers. Do these conform to the classical expectation that transaction prices will tend to adjust to equate price and marginal cost?

If one goes back of the wholesale price indexes published by the *BLS* to the specific price series of individual reporters there is ample evidence of the infrequency of price change and quantum jumps when prices do change.<sup>17</sup> However, the critics of the administered-price thesis have rejected this evidence as not reflecting true transaction prices. By obtaining data on prices paid by buyers, the authors of the *NB* report have obtained data which they vouch for as representing true transaction prices. Altogether they obtained more than 1200 individual price series from individual reporters. Unfortunately the individual series have not been published and their specific behavior tends to be lost in the compiled and published *NB* indexes.

Fortunately, in describing the methods used in compiling commodity indexes, the authors do present four of the thirteen or more price series underlying their index for a single commodity. This commodity is bulk ammonia. It was chosen by the authors as "... our illustrative commodity ..." (p. 66) and it is fair to assume that it is not seriously atypical, at least of administra-

tive prices. The four sample price series for ammonia "... were chosen to display the varieties of price data reported: unchanging prices, irregularly changing prices, broken price series, and frequently changing prices" (p. 39). The four series are presented in chart form, reproduced here as Figure 3. Since it is the only specific data on actual prices reported, it will be used here to sharpen the essence of the challenge to classical theory represented by the behavior of administered prices.

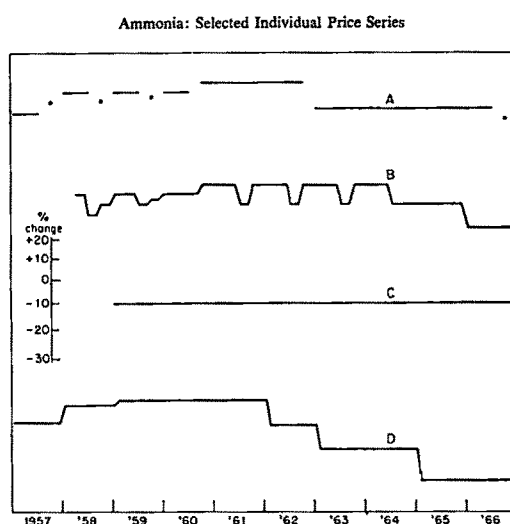


FIGURE 3

Examination of Figure 3 shows great infrequency of price change and large quantum jumps. Buyer C reported that it paid *exactly the same price* each month for eight years. Buyer A indicated that it paid only four different prices in ten years. Buyer B reported much more frequent price changes, fifteen times in nine years. However, all but three of these are seasonal and the administered-price thesis is not concerned with regular seasonal price jumps.<sup>18</sup> Leaving out this seasonal, the

<sup>17</sup> In part the infrequency and quantum jumps are reflected in the published indexes for specific commodities but in part they disappear in the smoothing effect of combining the raw data supplied by several reporters into a single published commodity index.

<sup>18</sup> Where a seasonal change in price is not exactly canceled by equal seasonal changes in the opposite direction within seven months, it is treated as a case of nonseasonal change.

four series taken together show eleven changes in a combined life of thirty-six years or an average of one change in every three and a quarter years. That these are "genuine transaction prices" paid by buyer *A* is asserted in the report and each of these series is clearly for an administration-dominated price, not a classically adjusting price.

Figure 3 also shows that when price change does occur it is likely to be unexpectedly large. The three changes in the price paid by buyer *A* averaged 8 percent with the largest around 11 percent.<sup>19</sup> For buyer *B* half of the seasonal changes amounted to around 8 percent and the three nonseasonal changes averaged 8 percent. For buyer *D* the five changes averaged 9 percent with the largest single month-to-month change a drop of 13 percent.

The infrequency of price change and the very large quantum jumps may represent a fairly extreme example of administered-price behavior, but this small sample of "genuine transaction prices" does serve to raise the crucial question: how can this price behavior be explained within the framework of traditional theory?

First, how can the differences in the behavior of the transaction prices paid by the four ammonia buyers be explained in terms of classical supply and demand theory? If the four buyers are buying essentially the same product in the same market, it would seem impossible even with the utmost stretching of classical theory to bring the behavior of the four different price series within its scope. If the four series are treated as involving four different products due to geographical or other forms of differentiation but generally related through the market, where is the evidence that there is a relation in

their price movements? And if they are treated as four independent products, how is it possible for them to be independent when they serve so largely the same uses? In denigrating the *BLS* series, the report says "... the *BLS* pattern is baffling—how can one seller ignore changes in the price of the same commodity made by rivals?" (p. 66). We can well ask the same question with respect to the buyers and sellers of the four ammonia price series.

Second, how can the long periods of constant prices be explained? Does this mean supply and demand conditions in the ammonia market from which buyer *C* obtained bulk ammonia were perfectly stable? Or even approximately stable? Were demand and marginal cost both constant throughout eight years? Or did both change in such a way as to just offset each other? Did the same price produce an equating of price and marginal cost for a period of eight years?

Third, how can the quantum jumps be explained? Is there likely to be no change in demand or costs for months of years at a time and then a sudden change requiring an 8 percent change in price?

Although the authors present this data for their "illustrative commodity," they reject a priori the evidence of price behavior which it carries. Their comparison of the *BLS* and *NB* price indexes for ammonia revealed a tendency for the *BLS* index to change infrequently, eighteen changes in ten years, with half the changes between 5 and 10 percent. In contrast, the *NB* index changed four times as frequently and the distribution of changes is bell-shaped with 90 percent of the changes less than 1 percent and only one change over 5 percent. Of this difference they say "The *NB* distribution is intuitively much more plausible than the *BLS* distribution, which would reflect a *world of alternating rigidity and fitful shifts in supply and demand conditions*" (p. 66, emphasis added).

<sup>19</sup> No numerical data is given for these four price series and the figures in this paragraph are read off of the published chart with some possible small error.

Yet if the four ammonia series are representative of the price behavior reported in all the ammonia series used in compiling the *NB* ammonia index, then the world from which the index is drawn is a "world of alternating rigidity and fitful shifts." Whether the rigidities and fitful shifts are in supply and demand conditions or only in prices is another matter.

The difference in the behavior of the two indexes is easily explained by the smoothing effect of a larger number of reporters and the peculiar effects of the *NB* procedure used in compiling the index. The *NB* ammonia index is compiled from price series supplied by thirteen or more reporters as compared to around three to five reporters for the *BLS* index.<sup>20</sup> The larger the number of reporters, the more frequent the changes in the combined index and the smaller the average change. Thus, if there were ten reporters and only one changed its price in a particular month and changed its price by 8 percent, the *combined* index would show a change of .8 of 1 percent. Similarly, if in a three-month period, three buyers each reported an 8 percent price increase but each reported in a different month and other buyers reported no change, the combined index would show three changes, each less than 1 percent.

Smoothing would also arise from the use of linear interpolation to fill the gaps in the series provided by a single reporter like those in the series reported by ammonia buyer *A*.<sup>21</sup> These gaps which could be as

much as fifteen months would provide as many artificial small changes as months in the gap where a single observed difference in price was interpolated. Even the combining of the four ammonia series into a single index with linear interpolation for *A* would result in showing thirty-six price changes in ten years when the average change per reporter was only six. It would also show an *average* price change of less than 1.5 percent instead of the actual 8 percent.

This artificial smoothing of the frequency and size of the actual price changes in the underlying data covers up the actual rigidity and quantum jumps which are characteristic of administered prices. Yet the authors cite Alfred Marshall's famous dictum that *natura non facit saltum* in support of their contention that the greater *smoothness* of the *NB* indexes as compared with the corresponding *BLS* indexes (based on fewer reporters) is evidence favorable to the greater validity of the *NB* indexes. (See S and K, p. 9.) In this day of world-wide knowledge of earthquakes, quantum mechanics and biological mutations it is somewhat quaint to offer the view that "nature does not like jumps." It is hardly a reason for not facing up to the implications of the price jumps in the actual data or the challenge to classical theory which they present.

## VII. Conclusion

The preceding analysis points to four major conclusions:

First, the new price data collected from buyers by the National Bureau strongly support the idea that, in business cycles, administration-dominated prices tend to behave quite differently from market-dominated prices and, more specifically,

<sup>20</sup> The *BLS* ammonia price series 0611-13 used by Stigler and Kindahl was a refrigerator type of anhydrous ammonia and was obtained from *Oil, Paint and Drug Reporter*. At the office of the *Reporter* the exact number of reporters from 1957 to 1966 was not known but was said to be "probably at least three and not more than five." Allied Chemical and duPont were specifically named and it was thought that there were probably two more.

<sup>21</sup> In some cases, linear interpolation alone was used. In others, linear interpolation was then modified by an index derived from other reports for the same commod-

ity. (See S and K, p. 103.) In either case the resultant filled-in price series for the single reporter would show small change in every month of the gap except by chance.

tend to fall less or not fall at all in cyclical recessions and to rise less or not rise at all in cyclical recoveries, thus confirming the administered-price thesis.

Second, the new price data give a new dimension to the administered-price thesis by disclosing a substantial number of prices which tend to rise with cyclical recessions and a substantial number which tend to fall with cyclical recovery.

Third, the new price data fail to support the main conclusion of the National Bureau report that there is a preponderant tendency for the industrial prices covered in the report to move in response to the cyclical movement of business. When market-dominated prices are excluded, the failure to conform to the classical expectation is even greater.

Fourth, the actual behavior of administration-dominated prices in the sample tends to differ so sharply from the behavior to be expected from classical theory as to challenge the basic conclusions of that theory. However well the theory may apply to market-dominated prices, it would not seem to apply to the bulk of the administration-dominated prices in the sample or to that part of the industrial world which they typify.

Until economic theory can explain and take into account the implications of this nonclassical behavior of administered prices, it provides a poor basis for public policy. The challenge which administered prices make to classical economics is as fundamental as that made by the quantum to classical physics.

## APPENDIX A

*Behavior of Fifty "Commodity" Indexes in Contraction and Recovery<sup>a</sup>*

Commodity	NB Index Number	1st Contraction	1st Recovery	2d Contraction	2d Recovery
Not Conforming to Classical Expectation in all Four Opportunities					
Phenol	49	<i>Up</i>	<i>Down</i>	<i>Up</i>	<i>Down</i>
Phenolic Resins	53	<i>Up</i>	<i>Down</i>	<i>Up</i>	<i>Down</i>
Ammonia	40	<i>Up</i>	<i>Down</i>	<i>Up</i>	<i>No Change</i>
Phthalic Anhydride	48	<i>No Change</i>	<i>Down</i>	<i>Up</i>	<i>Down</i>
Steel Sheet & Strip	1	<i>Up</i>	<i>No Change</i>	<i>No Change</i>	<i>No Change</i>
Stainless Sh. & St.	8	<i>No Change</i>	<i>No Change</i>	<i>No Change</i>	<i>Down</i>
Industrial Belting	28	<i>No Change</i>	<i>No Change</i>	<i>Up</i>	<i>No Change</i>
Paperboard	32	<i>No Change</i>	<i>Down</i>	<i>No Change</i>	<i>No Change</i>
Chlorine	38	<i>No Change</i>	<i>No Change</i>	<i>No Change</i>	<i>Down</i>
Neoprene	27	<i>No Change</i>	<i>No Change</i>	<i>No Change</i>	<i>No Change</i>
Not Conforming to Classical Expectation in Three of Four Opportunities					
Gasoline	20	<i>Down</i>	<i>Down</i>	<i>Up</i>	<i>Down</i>
Fuel Oil	22	<i>Down</i>	<i>Down</i>	<i>Up</i>	<i>Down</i>
Titanium Dioxide	37	<i>Up</i>	<i>Down</i>	<i>Down</i>	<i>Down</i>
Oxygen	39	<i>Up</i>	<i>Down</i>	<i>Up</i>	<i>Up</i>
Benzene	43	<i>Up</i>	<i>Down</i>	<i>Down</i>	<i>Down</i>
Styrene Monomer	44	<i>Down</i>	<i>Down</i>	<i>Up</i>	<i>Down</i>
Antibiotics	54	<i>Down</i>	<i>Down</i>	<i>Up</i>	<i>Down</i>
Steel Wire	7	<i>Up</i>	<i>No Change</i>	<i>Down</i>	<i>Down</i>
Acetylene	42	<i>Up</i>	<i>Up</i>	<i>Up</i>	<i>No Change</i>
Polyvinyl Chloride	52	<i>No Change</i>	<i>Down</i>	<i>Down</i>	<i>Down</i>
Tranquilizers	55	<i>No Change</i>	<i>Down</i>	<i>Down</i>	<i>Down</i>
Steel Bars & Rods	5	<i>Up</i>	<i>Up</i>	<i>No Change</i>	<i>No Change</i>
Sulphuric Acid	35	<i>No Change</i>	<i>No Change</i>	<i>Down</i>	<i>Down</i>
Acetone	41	<i>No Change</i>	<i>No Change</i>	<i>Down</i>	<i>Down</i>
Cardiac Glycosides	56	<i>No Change</i>	<i>Up</i>	<i>No Change</i>	<i>Down</i>
Steel Sheet & Strip, HR	2	<i>No Change</i>	<i>Up</i>	<i>No Change</i>	<i>No Change</i>
Steel Plates	4	<i>No Change</i>	<i>Up</i>	<i>No Change</i>	<i>No Change</i>
Alloy Steel Bars	9	<i>No Change</i>	<i>Up</i>	<i>No Change</i>	<i>No Change</i>
Newsprint	30	<i>No Change</i>	<i>No Change</i>	<i>No Change</i>	<i>Up</i>
Ethyl Alcohol	45	<i>No Change</i>	<i>Up</i>	<i>No Change</i>	<i>No Change</i>
Not Conforming to Classical Expectation in Two of Four Opportunities					
Methyl Alcohol	46	<i>Down</i>	<i>Down</i>	<i>Down</i>	<i>Down</i>
Polyethalene	50	<i>Down</i>	<i>Down</i>	<i>Down</i>	<i>Down</i>
Polystyrene	51	<i>Down</i>	<i>Down</i>	<i>Down</i>	<i>Down</i>
Plate Glass	59	<i>Up</i>	<i>Up</i>	<i>Down</i>	<i>Down</i>
Electric Motors	61	<i>Down</i>	<i>Down</i>	<i>Down</i>	<i>Down</i>
Carbon Steel Pipe	6	<i>Up</i>	<i>Up</i>	<i>Down</i>	<i>Down</i>
Truck and Bus Tires	25	<i>Up</i>	<i>Up</i>	<i>Down</i>	<i>Down</i>
Caustic Soda	36	<i>Down</i>	<i>Down</i>	<i>Down</i>	<i>Down</i>
Passenger Car Tires	24	<i>Up</i>	<i>Up</i>	<i>Down</i>	<i>No Change</i>
Synthetic Rubber	26	<i>No Change</i>	<i>Up</i>	<i>Down</i>	<i>Down</i>
Book & Magazine Paper	29	<i>Down</i>	<i>Up</i>	<i>No Change</i>	<i>Down</i>
Kraft Papers	31	<i>Down</i>	<i>Down</i>	<i>Down</i>	<i>No Change</i>
Paper Boxes	33	<i>Down</i>	<i>Down</i>	<i>Down</i>	<i>No Change</i>
Bond Paper	34	<i>No Change</i>	<i>Up</i>	<i>Down</i>	<i>Down</i>
Paint	57	<i>Down</i>	<i>No Change</i>	<i>Up</i>	<i>Up</i>
Portland Cement	58	<i>Up</i>	<i>Up</i>	<i>No Change</i>	<i>Up</i>
Safety & Window Glass	60	<i>Down</i>	<i>Down</i>	<i>Down</i>	<i>No Change</i>
Tinplate	3	<i>No Change</i>	<i>Up</i>	<i>Down</i>	<i>No Change</i>
Not Conforming to Classical Expectation in One of Four Opportunities					
Diesel & Dist. Fuel	21	<i>Down</i>	<i>Down</i>	<i>Down</i>	<i>Up</i>
Glycerine	47	<i>Down</i>	<i>Up</i>	<i>Down</i>	<i>Down</i>
Conforming to Classical Expectation in All Four Opportunities					
None	—	—	—	—	—

<sup>a</sup> Source: S and K, pp. 108-71. The 63 commodity indexes for which 10-year data is given less 13 market-dominated indexes. First contraction and recovery July 1957-April 1958-June 1959; second, January 1960-January 1961-March 1962. No change includes all indexes changing less than  $\pm .05$  percent per month between respective dates. Italicized observations failed to move in conformity with changes in industrial activity.

*Commodities Classed as Market-Dominated*

NB Index Number	Commodity	Reason
C-23	Bituminous Coal	Competitive market
C-63	Softwood Plywood	Competitive market
C-64	Wood Flooring	Competitive market
C-10	Aluminum Ingot and Shot	World Market
C-13	Copper Ingot	World Market
C-11	Aluminum Sheet and Strip	Raw Material Dominated
C-12	Aluminum Wire and Cable	Raw Material Dominated
C-14	Copper Pipe and Tubing	Raw Material Dominated
No data	Copper Sheet and Strip	Raw Material Dominated
C-15	Copper Wire and Cable	Raw Material Dominated
C-16	Insulated Copper Wire	Raw Material Dominated
C-17	Copper Magnet Wire	Raw Material Dominated
C-18	Zinc Products	Raw Material Dominated
C-19	Brass Bars and Rods	Raw Material Dominated
No data	Brass Sheet and Strip	Raw Material Dominated

## APPENDIX B

*Market-Dominated Prices*

While the sample of industrial prices covered in the study purports to have been chosen from those which have figured prominently in the discussion of administered prices, it is clear that it includes some market-dominated prices. Just which items belong in this category can be a matter of debate since market power where it exists is itself a matter of degree. There can be little question, however, that Plywood (softwood) and Wood Flooring are sold under highly competitive conditions with the largest four producers accounting for only 18 and 17 percent, respectively, of the value of shipments in 1958.<sup>22</sup> Copper, though its refining and fabrication in the United States is highly concentrated, is competitively priced in the London Metals Exchange (*LME*) for the world market and the U.S. price, at the time of the two contractions, was a few cents over the *LME* price but fluctuated with it.<sup>23</sup> Also the prices of simply fabricated copper products like copper wire and copper tubing are so dominated by the price of copper that

they are properly classed as market-dominated prices. Altogether, at least fifteen of sixty-eight commodities should be classed as having market-dominated prices, three because of low concentration, two because of competitive world markets and ten because the price was dominated by the flexible market price of its raw material. The list is shown above.

## REFERENCES

- G. C. Means, *Industrial Prices and their Relative Inflexibility*, U.S. Senate Document 13, 74th Congress, 1st Session, Washington 1935.
- *The Structure of the American Economy, Part I, Basic Characteristics*, National Resources Committee, Washington 1939.
- *The Corporate Revolution in America*, New York 1954, pp. 101-17.
- *Hearings on Administered Prices, Part 9*, Senate Subcommittee on Antitrust and Monopoly, Washington 1959, pp. 4745-60.
- G. J. Stigler, "Editor's Note," *J. Econ. Lit.*, Sept. 1971, 9, 852.
- and J. K. Kindahl, *The Behavior of Industrial Prices*, New York 1970.
- U.S. Dept. of Commerce, *Concentration Ratios in Manufacturing Industry*, Washington 1962.

<sup>22</sup> See U.S. Dept. of Commerce pp. 112 and 122.

<sup>23</sup> At some time in 1964 or 1965, a complicated two-price system was adopted in the United States but this was well after the two contractions involved in the NB study.

# On Taxation and the Control of Externalities

By WILLIAM J. BAUMOL\*

It is ironic that just at the moment when the Pigouvian tradition has some hope of acceptance in application it should find itself under a cloud in the theoretical literature. James Buchanan has argued that its recommended taxes and subsidies may even increase resource misallocation in the presence of monopoly. Otto Davis and Andrew Whinston (1962) have, in effect, raised doubts about its applicability in the presence of oligopoly. And Ronald Coase has asserted that the tradition has not selected the correct taxation principle for the elimination of externalities, and may not even have chosen the right individuals to tax or to subsidize. In this paper I will suggest that these authors have led the discussion in our profession to focus on the wrong difficulties. In doing so they have, albeit inadvertently, drawn attention away from some of the most important limitations of the Pigouvian prescription as an instrument of policy and from con-

sideration of the means that might prove effective in practice.

The main purpose of the paper is to show that, taken on its own grounds, the conclusions of the Pigouvian tradition are, in fact, impeccable. Despite the various criticisms that have been raised against it in the large numbers case, which is of primary importance in reality and to which Pigou's analysis directs itself, his tax-subsidy programs are generally those required for an optimal allocation of resources. Moreover, I will attempt to show that where an externality is (like the usual pollution problem) of the public goods variety, neither compensation to nor taxation of those who are affected by it is compatible with optimal resource allocation. Pigouvian taxes (subsidies) upon the generator of the externality are all that is required.

However, as is well known, the Pigouvian proposals suffer from a number of serious shortcomings as operational criteria when one seeks to implement them precisely as they emerge from the theory. I therefore discuss a modified approach that recommends itself more for its promise of effectiveness, than its theoretical nicety. It consists of two basic steps: the setting of standards, more or less arbitrarily, of levels of pollution, congestion and the like, that are considered to be tolerable, and the design of taxes and effluent charges whose rates are shown by experience to be sufficient to achieve the selected standards of acceptability. Such a system of charges will, at least in principle, effect any preselected reduction in,

\* Professor of economics, Princeton University and New York University. I would like to express my gratitude to the National Science Foundation whose assistance helped materially in the completion of the paper and to my colleagues James Litvack, Wallace Oates, and David Bradford, to my students Mark Gaudry and Bryan Boulter, and to Peter Bohm, James Buchanan, Ronald Coase, Karl-Göran Mäler, Herbert Mohring, and Ralph Turvey who have given me many very helpful suggestions, and saved me from a number of serious errors. Mohring and J. Hayden Boyd have written an extremely illuminating paper dealing, among other relevant matters, with the portions of the Coase-Buchanan-Turvey arguments in the case where the polluters and their victims "can and do negotiate." Since the present paper concerns itself only with the "relevant" large numbers case where there is no negotiation, it deliberately makes no attempt to consider the interesting negotiation case examined so helpfully by Mohring and Boyd.

say, the pollution content of our rivers, at minimum cost to society. It automatically achieves an efficient allocation of the required reduction in emissions among the offending firms *even if they are neither pure competitors nor profit maximizers*. Thus, a persuasive case can be made for the use of taxes and subsidies to control externalities, even if they will not produce an optimal allocation of resources in the complex world of reality.

### I. The Coase Argument in the Case Without Negotiation

Recommendations designed for the competitive case can clearly run into difficulties in the presence of monopolistic elements. Buchanan reminds us that, if a polluting monopolistic industry already restricts the outputs of its products below their competitive levels, the imposition of an effluent charge to restrict output still further is hardly likely to be appropriate. And Davis and Whinston (1962) show for the case of externalities under oligopoly that it is rather difficult to come up with an ideal set of taxes since in the small numbers case just about anything is possible by way of pricing and output levels. However, these arguments have little direct bearing on the Pigouvian analysis because it is couched entirely in terms of pure competition (on this see Stanislaw Welicz' illuminating discussion), which, in view of the large numbers involved in virtually all of the externalities problems that worry us today, is entirely apropos.

Coase's arguments, buttressed by impressive legal erudition, are less easily dealt with. He offers us a number of illuminating observations, among them the interesting point (see his Section IV) that (in the relatively unimportant cases) where only a small number of decision makers is involved, a process of voluntary bargaining and side payments among those con-

cerned by an externality may produce an optimal allocation of resources, even in the absence of liability for damage. This implies that where small numbers are involved, the imposition of a "corrective" Pigouvian tax may be too much of a good thing—it can produce a misallocation rather than eliminating it.

Coase suggests, however, that even in cases where there is no negotiation among the parties affected by an externality the Pigouvian taxes and subsidies may be the wrong remedy—that they may only modify the character of the misallocation of resources. Coase's central argument appears to be the following: Every social cost is inherently reciprocal in nature. The nearby residents who breathe smoke spewn by a factory must share with the management of the factory the responsibility for the resulting social cost. True, if the factory were closed up the social cost would disappear. But the same holds for its neighbors—were they to move away no one would suffer smoke nuisance. Put another way, just as the smoke emitted by the factory imposes at least a psychic cost on its neighbors, the latter's insistence on the installation of purification devices or a reduction in the pollution-producing activity imposes a cost on the factory.

This position, though at first glance very odd (the murder victim too, is then always an accessory to the crime), grows more persuasive as one considers it further. Coase does not raise the issue as a matter of distributive justice. Rather, he suggests, because of the reciprocal structure of the externality, the traditional taxes and subsidies are likely to lead to a misallocation of resources.<sup>1</sup> If it is socially less costly to

<sup>1</sup> Thus Coase starts out with

... the case of a confectioner, the noise and vibrations from whose machinery disturbed a doctor in his work. To avoid harming the doctor would inflict harm on [be costly to] the confectioner. The problem posed by this case was essentially whether it was worthwhile, as a

remove the neighbors from the vicinity of the factory than to reduce the quantity of pollutants emitted by the plant (taking into account the location preferences of the current residents), surely the former is the course of action which is more desirable socially.

In that case, should not a tax sometimes be levied, at least in part on those who choose to live near the factory rather than upon the factory owners?<sup>2</sup> Otherwise might not too many persons be induced to move near the factory thus, incidentally, increasing the magnitude of the Pigouvian tax since the social damage caused by the smoke must then rise correspondingly?

A simple model shows readily that, properly stated, the prescription of the Pigouvian tradition is (at least formally) correct. An appropriately chosen tax, levied only on the factory (without payment of

---

result of restricting the methods of production which could be used by the confectioner, to secure more doctoring at the cost of a reduced supply of confectionery products. [Section II, p. 2]

<sup>2</sup> If the factory owner is to be made to pay a tax equal to the damage caused, it would clearly be desirable to institute a double tax system and to make residents of the district pay an amount equal to the additional cost incurred by the factory owner (or the consumers of his products) in order to avoid the damage. [Coase, Section IX, p. 41] An even stronger statement on this subject occurs in Buchanan and Stubblebine (Section III):

... full Pareto equilibrium can *never* be attained via the imposition of unilaterally imposed taxes and subsidies until all marginal externalities are eliminated. If a tax subsidy method, rather than 'trade,' is to be introduced, it should involve bi-lateral taxes (subsidies). Not only must *B*'s behavior be modified so as to insure that he will take the costs externally imposed on *A* into account, but *A*'s behavior must be modified so as to insure that he will take the costs 'internally' imposed on *B* into account. [italics added]

However, in a recent letter Buchanan commented:

In my own thinking . . . I did not ever think of this sort of [double] tax at all, and it would have surely seemed bizarre to me to suggest that taxes be levied on both the factory and the laundries. What we were proposing was the Wicksellian public-goods approach. Suppose that existing property rights allow the factory to put out the smoke . . . There is a public goods problem here; the residents get together, impose a tax on *themselves* to subsidize the factory to install the smoke prevention device.

compensation to local residents) is precisely what is needed for optimal resource allocation under pure competition. No tax on nearby residents is required or, taken in real terms, is even compatible with optimal resource allocation. Thus the obvious and apparently common interpretation of the Coase position is simply invalid. We will see, however, that the issue Coase himself intended to raise was rather more subtle and his conclusions are not necessarily at variance with the Pigouvian prescription as I interpret it.

## II. Analysis: Should the Victims of Externalities be Taxed or Compensated?

To formalize the argument we construct an elementary general equilibrium model designed to represent in most explicit form the conditions envisioned in the Coase argument, departing from it only by an assumption of universal perfect competition, including thereby the critical stipulation that costs of negotiated and voluntary control of externalities are prohibitive. In addition, we adopt the simplifying premises that there is only one scarce resource, labor, and that the externality (smoke) only affects the cost of production of neighboring laundries, rather than causing disutility for consumers. It is easy to show (see for example, fn. 5) that neither of these simplifications, nor the assumption that there are only four activities, affects the substance of the discussion. We utilize the following notation: Let

$x_1, x_2, x_3$ , and  $x_4$  be the outputs of the economy's four activities, I, II, III, and IV

$R$  be the total supply of the labor resource available

$x_5$  be the unused quantity of labor (which is assumed to be utilized as leisure)

$x_{ij}$  be the quantity of  $x_i$  consumed by individual  $j$  ( $i=1, \dots, 5$ ) ( $j=1, \dots, m$ )

$p_1, p_2, p_3, p_4$ , and  $p_5$  be the prices of the four outputs and leisure

$u_j(x_{1j}, \dots, x_{5j})$  be the utility function of individual  $j$ , and

$c_1(x_1)$ ,  $c_2(x_1, x_2)$ ,  $c_3(x_3)$  and  $c_4(x_4)$  be the respective total labor cost functions for our four outputs

Here  $x_1$  is an output whose production imposes external costs on the manufacture of  $x_2$  (say, industry II is the oft-cited laundry industry whose costs are increased by I's smoke). To permit the full range of Coase's alternatives (moving of the factory's neighbors and elimination of smoke by the factory), each of these two products is taken to have a perfect substitute. The substitute for  $x_1$  is  $x_3$  whose production yields no externalities, but whose cost is different (presumably higher) than that of  $x_1$ . We may think of commodity III as identical with I, but produced in a factory equipped with smoke elimination equipment. Similarly, industry IV is taken to offer the same output as II but its operations have been relocated (at a cost) in order to avoid the effects of the externalities.<sup>3</sup> Thus, by changing the ratio between  $x_2$  and  $x_4$  the model can relocate as much of the laundry output as is desired.

All prices are expressed in terms of hours of labor so that, identically,

$$(1) \quad p_5 = 1.$$

<sup>3</sup> Since product III is a perfect substitute for product I and product IV is a perfect substitute for product II, the utility function for individual  $j$  can be written as  $u_j(x_{1j}+x_{3j}, x_{2j}+x_{4j}, x_{5j})$ . This is, of course, a special case of the more general utility function utilized in the text, and as the reader can verify, the conclusions are totally unaffected by the use of the particular form of the utility function just described.

Pareto optimality then requires maximization of the utility of any arbitrarily chosen individual, say  $m$ , subject to the requirement that there be no loss in utility to any of the  $m-1$  other persons, i.e., given any feasible level for these other persons' utility. Thus the problem is<sup>4</sup> to maximize

$$u_m(x_{1m}, \dots, x_{5m})$$

subject to

$$u_j(x_{1j}, \dots, x_{5j}) = k_j \text{ (constant)} \\ (j = 1, 2, \dots, m-1)$$

$$\sum_{j=1}^m x_{ij} = x_i \quad (i = 1, \dots, 5)$$

and the labor requirement (production function) constraint

$$c_1(x_1) + c_2(x_1, x_2) + c_3(x_3) + c_4(x_4) + x_5 = R$$

We immediately obtain our Lagrangian

$$(2) \quad L = \sum_{j=1}^m \lambda_j [u_j(x_{1j}, \dots, x_{5j}) - k_j] \\ + \sum_i \nu_i (x_i - \sum_j x_{ij}) \\ + \mu [R - c_1(x_1) - c_2(x_1, x_2) \\ - c_3(x_3) - c_4(x_4) - x_5]$$

where we may take  $\lambda_m = 1$ ,  $k_m = 0$ .

We use the notation  $u_{ji}$  to represent  $\partial u_j / \partial x_{ij}$  and  $c_{ik}$  to represent  $\partial c_i / \partial x_k$  (or  $dc_i / dx_k$ , where appropriate).

Then, differentiating in turn with respect to the  $x_{ij}$  and the  $x_i$  we obtain the first-order conditions

$$\partial L / \partial x_{ij} = \lambda_j u_{ji} - \nu_i = 0 \quad (i = 1, \dots, 5) \\ (j = 1, \dots, m)$$

$$\partial L / \partial x_1 = -\mu(c_{11} + c_{21}) + \nu_1 = 0$$

$$\partial L / \partial x_i = -\mu c_{ii} + \nu_i = 0 \quad (i = 2, 3, 4)$$

$$\partial L / \partial x_5 = -\mu + \nu_5 = 0$$

<sup>4</sup> For a more sophisticated variant of this model, using the techniques of non-linear programming, see Robert Meyer.

Now, from consumer equilibrium analysis, we know that for any two commodities,  $a$  and  $b$ , and any two prices,  $p_a$  and  $p_b$ , we have  $p_a/p_b = u_{ja}/u_{jb}$  ( $j=1, \dots, m$ ) or  $\omega_j p_i = u_{ji}$  for all  $i$  and some  $\omega_j$ .

Hence,  $\lambda_j u_{ji} = \lambda_j \omega_j p_i$ , so that writing  $s_j = \lambda_j \omega_j$  the first of our first-order conditions becomes  $v_i = s_j p_i$  for all individuals,  $j$ . Consequently the value of  $s_j$  must equal the same number,  $s = v_i/p_i$  for every individual, and that first equation of the first-order conditions now becomes simply  $v_i = s p_i$  for all  $i$ . Substituting this expression for  $v_i$  into the other first-order conditions, we obtain

$$s p_1 = \mu(c_{11} + c_{21})$$

$$s p_i = \mu c_{ii} \quad (i = 2, 3, 4)$$

$$(3) \quad s p_5 = s = \mu \quad \text{since } p_5 = 1 \text{ [by (1)]}$$

By (3) we may then divide through the preceding conditions by  $s = \mu$ , and they therefore reduce just to<sup>6</sup>

$$p_1 = c_{11} + c_{21}$$

$$p_2 = c_{22}$$

$$(4) \quad p_3 = c_{33}$$

$$p_4 = c_{44}$$

$$p_5 = 1$$

In other words, the optimal price for the externality-generating product is equal to the (Pareto optimal) level of its entire

<sup>6</sup> The analysis can also take account of constraints on the availability of land at the relevant locations, which give rise to rents that equalize costs at all locations actually utilized. If  $S_a$  and  $S_b$  represent the availability of land near and away from the factory, respectively, presumably we would add to the labor constraint in the model the two additional land-use constraints  $g_a(x_1, x_2, x_3) + s_a = S_a$  and  $g_b(x_4) + s_b = S_b$ , with the quantities of unused land,  $s_a$  and  $s_b$ , perhaps entering the utility functions. It then follows, just as before, that the equilibrium conditions are now  $p_1 = c_{11} + c_{21} + p_a g_{a1}$ ;  $p_2 = c_{22} + p_a g_{a2}$ ;  $p_3 = c_{33} + p_a g_{a3}$ ;  $p_4 = c_{44} + p_b g_{b4}$ ;  $p_5 = 1$ ;  $p_a = \rho_a/\mu$ ;  $p_b = \rho_b/\mu$ ; where  $\rho_a$  and  $\rho_b$  are the Lagrange multipliers for the new constraints and  $p_a$  and  $p_b$  are the (labor) prices of land at the two locations. Our previous conclusions are, thus, totally unaffected. Only the smoke producer's product sells for more than its marginal private cost of labor plus land.

social<sup>6</sup> marginal cost,  $c_{11} + c_{21}$ , while the optimal price for any item,  $i$ , which generates no externalities is simply its marginal private cost,  $c_{ii}$ . To obtain these prices in our world of pure competition, one need merely levy an excise tax on item 1 equal to  $c_{21}$  (labor hours) dollars per unit, just as the Pigouvian tradition requires. Assuming the appropriate concavity-convexity conditions hold, this will automatically satisfy the necessary and sufficient conditions for the Pareto optimal output levels.<sup>7</sup> In the competitive case, where negotiation is impractical, that is all there is to the matter. The generalization to the case of  $n$  outputs, each of them imposing externalities on a number of the others, is immediate.

It is important to observe that, *the solution calls for neither taxes upon  $x_2$ , the neighboring laundry output, nor compensation to that industry for the damage it suffers.*

One way to look at the reason is that our model (and the pollution model in general) refers to the important case of *public* externalities. The laundry whose output is

<sup>6</sup> The social cost is not  $c_{21}$  alone but is the sum of the private and the external costs together (see the illuminating terminological discussion by D. W. Pearce and Stanley Sturmev). Note that the tax, implicitly, is a tax on *smoke* not a tax on  $x_1$ , the output of the smoke producing industry. For if  $s$  is the quantity of smoke and  $t$  the unit tax we may write  $t = c_{21} = (\partial c_2 / \partial s)(ds/dx_1)$  and obviously the firm can reduce its tax rate by decreasing the second of these terms, the smokiness of its product. This point has been emphasized by Charles Plott, who showed that a fixed tax per unit of  $x_1$  might even conceivably increase  $s$ , if  $s$  were an inferior input.

<sup>7</sup> Moreover, measured in real terms this is the only tax arrangement that satisfies the optimality requirements, neglecting the possibility of a lump sum tax or subsidy which does not affect the marginal conditions. F. Trenery Dolbear has shown that it is generally not possible to find an optimal tax rate that compensates fully those who suffer the effects of the externality. Since no compensation is paid to industry II, the solution that is derived here does not run into Dolbear's problem. We also do not run into the problem of a multiplicity of solutions corresponding to the various points on Dolbear's contract curve because we are dealing with a world of pure competition with a given initial distribution.

damaged by smoky air does not, by an increase in its own output, make the air cleaner or dirtier for others. As with all public goods, an increase in one user's consumption does not reduce the available supply to others.<sup>8</sup> Hence, the appropriate price (compensation) to a user of a public good (victim of a public externality) is *zero* except, of course, for lump sum payments. Thus, perhaps, rather than saying there is no price that will yield an optimal quantity of a public good (externality), it may be more illuminating to say that a double price is required: a nonzero price (tax) to the supplier of the good, and a zero price to the consumer. Of course, no ordinary price can do this job, but a Pigouvian tax, without compensation to those affected by an externality, can indeed do the trick.

### III. What Prevents an Excessive Influx of Neighbors?

When only smoke emission is taxed, with the tax level based on the magnitude of  $x_2$ , nearby laundry output, what will prevent too many laundries from moving

<sup>8</sup> In his discussion of these matters Coase seems at one point to skate awfully close to an error analogous to the confusion between pecuniary and technological externalities. He writes (section IX):

The tax that would be imposed would . . . increase with an increase in the number of those in the vicinity . . . But people deciding to establish themselves in the vicinity of the factory will not take into account [the resulting] fall in the value of production which results from their presence. This failure to take into account costs imposed on others is comparable to the action of a factory-owner in not taking account the harm resulting from his emission of smoke. [p. 42]

This is analogous to the argument that where the supply curve of labor is rising an increase in output by firm *A* must produce externalities, by raising *B*'s labor costs. But, of course, this merely represents a transfer from *B* to his workers and is not a real net cost to society. For that reason, as is well known, pecuniary externalities do *not* lead to resource misallocation. Like a price change, the variation in taxes constitutes a pecuniary externality. Both have real consequences but they are merely "movements along" the production and utility functions, i.e., any given vector of inputs will be able to produce the same outputs as before the change in tax rates, and any vector of output levels will still be able to yield the same utility levels.

near the smoky factory? The answer is that, when the tax on the externality producer is set properly, the externalities themselves keep down the size of the nearby population. Moreover, the level of the tax will control both the magnitude of smoke emission and thereby (indirectly), the size of the nearby population. A high tax rate will discourage smoke and hence encourage migration into the neighborhood. A low tax rate will encourage smoke and, hence, drive residents away. A tax on smoke alone is all that is needed to control the magnitudes of *both* variables. That is why, as shown by the mathematics of the preceding section, just a tax on the smoke producer is sufficient to produce an optimal allocation of resources among all the activities in our model.<sup>9</sup>

A diagram may help to make the point clearer. Figure 1 shows the response of our two industries' outputs to a change in the tax rate on the polluting industry, I. We see that as the tax rate varies, industry I's output response follows the curve  $RR'$ . Thus, if the tax level is  $t$ , the output of industry I will be  $x_{1t}$ . But, because of the externalities, the output of industry II, in turn, reacts to the output of I. This relationship is described by reaction curve  $PP'$ . With  $x_1 = x_{1t}$  we see that  $x_2 = x_{2t}$ .

The tax rate on II can vary all the way from  $t=0$ , yielding output combination  $(x_{10}, x_{20})$ , to a prohibitive tax rate,  $t_p$ , that drives I out of business altogether, so that  $x_1=0$  and  $x_2=x_{2p}$ . Obviously, the ratio  $x_1/x_2$  then decreases monotonically as the tax rate increases and, assuming continuity, there will be some intermediate tax rate at which the two activities will be in balance. The tax will keep  $x_1$  in check while the external cost imposed by  $x_1$  on industry II will keep  $x_2$  to the right relative level. There is no need for a separate tax on II to achieve this goal.

<sup>9</sup> See the Appendix for a discussion of an argument by Buchanan and Stubblebine which is related to Coase's.

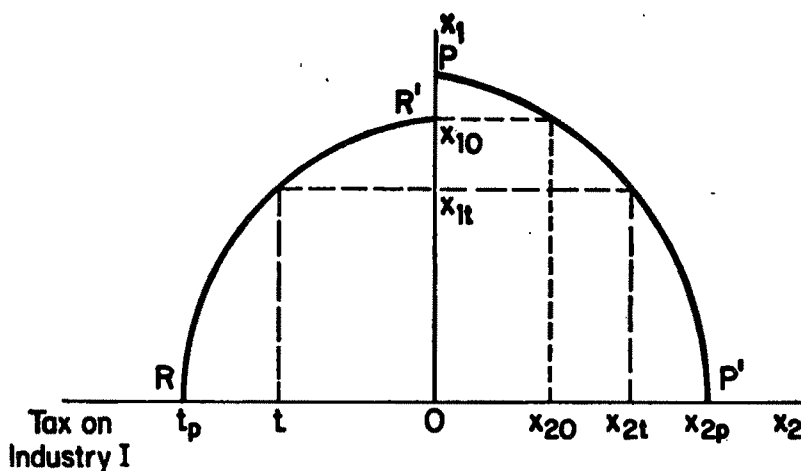


FIGURE 1

In order for this arrangement to work it is clearly necessary that the laundries *not* be compensated (at the margin) for the smoke damage they suffer. If they received in compensation an amount which varied with the magnitude of the smoke damage, that externality would not restrict the level of laundry activity near the factory. If the laundry operators' smoke costs were offset by damage compensation payments, obviously they would lose the economic incentive to eschew the vicinity of the smoky factory<sup>10</sup> and then Coase's tax on laundries would indeed be required to keep them away. But then the tax would be needed only to sop up the compensation payments which should never have been given in the first place.

#### IV. Multiple Local Maxima in the Coase Model

Coase's discussion is, however, right in pointing out the possibility that the econ-

omy may make the wrong choice between smoke elimination and laundry relocation: however the source of the problem, a multiplicity of local maxima, does not emerge clearly. Coase writes:

Assume that a factory which emits smoke is set up in a district previously free from smoke pollution, causing damage valued at \$100 per annum. Assume that the taxation solution is adopted and that the factory owner is taxed \$100 per annum as long as the factory emits the smoke. Assume further that a smoke-preventing device costing \$90 per annum to run is available. In these circumstances, the smoke-preventing device would be installed.

... Yet the position achieved may not be optimal. Suppose that those who suffer the damage could avoid it by moving to other locations or by taking various precautions which would cost them, or be equivalent to a loss in income of, \$40 per annum. Then there would be a gain in the value of production of \$50 if the factory continued to emit its smoke and those now in the district moved elsewhere or made other adjustments to avoid the damage. [Section IX]

One curious feature of this example is its assumption that while smoke damage is \$100, the cost of moving to other locations is only \$40. Under these circumstances one

<sup>10</sup> Of course, as smoke cost increases in the neighborhood of the factory, rents will fall to some extent and serve as partial compensation to the laundries. However, this does not change the analysis fundamentally. It is analogous to the case of rise in the price of an input which, as is well known, will tend to reduce the output of competitive firms, even though prices of other complementary inputs fall as a result. As the discussion of footnote 5 shows, explicit consideration of the price of land does not change the character of the solution.

may well wonder why people living near the factory do not just move elsewhere on their own initiative. Moreover, this may not simply be a matter of the numbers he happens to have chosen. The problem arises whenever the cost of moving away from the factory is less than the cost of elimination of the smoke, which in turn is less than the cost of the smoke damage, as the logic of Coase's example requires.

It is perhaps more important to recognize that the example presents us with a choice between (at least) two local optima. As will be argued later, a multiplicity of maxima is generally rendered more likely by the presence of externalities so that this issue is not a peculiarity of Coase's illustrations. The first of the two local optima in Coase's example (call it solution *A*) involves zero smoke emission and a full complement of residents near the factory. In the second optimum (solution *B*) no one remains in residence next to the factory and there is no restriction in smoke emission by the plant. Assuming that the (undesirable) initial position is the only other possibility, as Coase seems to suggest, which of these two will in fact be the global optimum depends on the cost of moving everyone away ( $m$  dollars) and the cost of elimination of the smoke ( $s$  dollars).

Assume with Coase that the initial cost of smoke damage is \$100, that  $s < 100$ , but that  $s < m$  so that it is cheaper to eliminate the smoke than to move the factory's neighbors. In this case, *A* is obviously the optimal solution. Since inhabitants surround the plant, and smoke emission, by assumption, cannot be changed by small amounts, the incremental social damage of an increase in smoke emission is \$100. Thus the correct Pigouvian tax is \$100 and, since  $s < 100$ , with such a tax it will pay the factory to do the right thing by society—to install the smoke eliminator.

Now assume instead that  $m < s < 100$  (it is cheaper to move people than to stop the smoke). This time *B* is the optimal solution, and since under *B* no one lives near the factory, the incremental cost of smoke is clearly zero. Therefore the proper Pigouvian tax is zero, a value that induces the factory to continue smoking, and its neighbors will find it advantageous (since  $100 > m$ ) to exit (coughing) from the area. Thus the zero Pigouvian tax value automatically satisfies the requirements of solution *B* when *B* is optimal just as the \$100 Pigouvian tax leads to solution *A* when *A* is optimal.

Of course, if *B* happens to be the true global optimum and society mistakenly imposes the \$100 Pigouvian tax appropriate for (local) optimum *A*, the economy may well end up with the inferior equilibrium *A*. This is the usual difficulty one encounters whenever there is a multiplicity of maxima, a problem that Pigou so clearly recognized (pp. 140, 224).

#### V. Departures from the Optimum and Adjustments in the Tax

If there is a departure from the optimal solution, for whatever reason, the value of the Pigouvian tax need not change. If, for example, *B* is the global optimum so that the optimal tax is zero, that tax need not be increased if a few (misguided) individuals choose to move back near the factory so that additional smoke now incurs (say) \$50 in damage. *At the optimal solution* the marginal cost of smoke is zero, and the equilibrium Pigouvian tax remains zero—it does not increase to \$50.

Here we have arrived at the issue which, I now understand, was really Coase's main point in the portion of his article we are considering. He writes in a letter:

... Let us assume your optimum tax is imposed. Now suppose that *A* establishes himself near the plant which produces the damaging emissions and thus

increases the amount of damage. Would your tax increase? My guess is that it would not (certainly if your tax system is right it should not). The tax system I was attacking was one which would in these circumstances, automatically lead to an increase in the tax as the damage increased.

This point is, surely, quite different from the issue he is usually interpreted to have raised (see the quotations in fn. 1, above, which suggest how the "usual interpretation" arose). It is, however, not inconsistent with the optimal solution derived in the previous section nor is it inconsistent with what I take to be the Pigouvian tradition.

But even on this issue Coase's strictures are not necessarily valid. Suppose that a regulator, having no way of calculating the *optimal* values of the Pigouvian tax is, however, able to determine the value of any marginal social damage at any point in time. *Faut de mieux* he therefore sets a tax rate equal to *current* marginal social damage on the smoke producer. This causes him to reduce his smoke, and so brings more laundries into the neighborhood. The tax is then readjusted to equal the new (higher) value of damage per puff of smoke, more laundries move in, and so on. Will this process of trial and error adjustments of the tax level, always setting it equal to current marginal smoke damage, converge to the optimum of Section II? That is, will the sequence of tax values converge to the optimal Pigouvian tax level, and will resource allocation approach optimality? That now seems to be Coase's main question.

Obviously, such a learning process always involves wastes and irreversibilities, just like the process of convergence of competitive prices to their equilibrium values in the absence of externalities. But if we follow the usual practice of assuming away these costs, one can show that the

process may be expected to converge to the optimum, provided the equilibrium is unique and stable. That is, there is then nothing inherently different about gradually moving taxes and prices towards their equilibrium here, and the process of adjustment toward competitive equilibrium when there are no externalities.

Specifically, letting  $s_t$  represent the tax per unit on commodity 1 at time  $t$ , and  $G_i$  be the  $i$ th adjustment function we may set

$$dx_{1t}/dt = G_1[p_{1t} - s_t - c_{11}(x_{1t})]$$

$$(5) \quad dx_{2t}/dt = G_2[p_{2t} - c_{22}(x_{1t}, x_{2t})]$$

$$dx_{it}/dt = G_i[p_{it} - c_{ii}(x_{it})] \quad (i = 3, 4)$$

$$(6) \quad s_t = c_{21}(x_{1t}, x_{2t}) \quad p_{it} = f_i(x_{1t}, \dots, x_{5t})$$

and where, as usual, we take

$$(7) \quad G_i(0) = 0$$

$$(8) \quad G'_i > 0$$

Going back to Section II, when optimality conditions (4) hold, we see by substituting them into (5) that all  $dx_{it}/dt = 0$ , i.e., (4) is indeed an equilibrium position for the dynamic system (5)–(8). Furthermore, any solution that does not satisfy (4) must involve at least one non-zero argument in the adjustment functions (5), and so no solution that fails to satisfy (4) can be an equilibrium.

It follows that if the dynamic system (5)–(8) is stable, and the solution to (4) is unique, the process with taxes set equal to *current* marginal damage and imposed *only on the polluter* will converge toward the optimum. One does not need to have calculated the optimal tax values from the beginning and stick to them.

The reason this process of simultaneous learning and adjustment does not work in Coase's example is that it involves (at least) two local maxima, as we have already noted. And in such a case, obviously, the adjustment mechanism may

well take us to the wrong maximum. Unfortunately, as we will see presently, in the presence of externalities, a multiplicity of maxima is all too likely to be with us.

## VI. Implementation Problems

Despite the validity in principle of the tax-subsidy approach of the Pigouvian tradition, in practice it suffers from serious difficulties. For we do not know how to estimate the magnitudes of the social costs, the data needed to implement the Pigouvian tax-subsidy proposals. For example, a very substantial portion of the cost of pollution is psychic; and even if we knew how to evaluate the psychic cost to some one individual we seem to have little hope of dealing with effects so widely diffused through the population.<sup>11</sup>

This would not necessarily be very serious if one could hope to learn by experience. One might try any plausible set of taxes and subsidies and then attempt, by a set of trial and error steps, to approach the desired magnitudes. Unfortunately, convergence toward the desired solution by an iterative procedure of this sort requires some sort of measure of the improvement (if any) that has been achieved at each step so that the next trial step can be adjusted accordingly. But we do not know the socially optimal composition of outputs, so we simply have no way of judging whether a given change in the trial tax values will even have moved matters in the right direction.

<sup>11</sup> For an excellent discussion of some of the work done in trying to implement Pigouvian taxes in practice, see Allen Kneese and Blair Bower, esp. ch. 6 and 8. The difficulty of determining the magnitude of the Pigouvian tax-subsidy level is one of Coase's major points, one that seems often to be overlooked in discussions of his paper. Thus Coase writes in a letter, "The view I expressed in my article was not that such an optimum tax system (levied solely on the damage producing firm) was inconceivable but that I could not see how the data on which it would have to be based could be assembled." An interesting approach to application for the small numbers case that is based on the decomposition principle of mathematical programming is presented by Davis and Whinston (1966).

These difficulties are compounded by another characteristic of externalities which has already been mentioned—the likelihood that in the presence of externalities there will be a multiplicity of local maxima (see Richard Portes, D. A. Starrett, and Baumol). Consequently, even if an iterative process were possible it might only drive us toward a local maximum, and may thus fail to take advantage of the really significant opportunities to improve economic welfare.

A simple model in the spirit of that of Section II can be used to show that the presence of "strong" externalities can be expected to produce a violation of the convexity conditions in whose absence one normally finds a multiplicity of local optima.

Let us assume (to permit the use of a two-dimensional diagram) that there exist only the first two of our four activities (the smoky output,  $x_1$ , and nearby laundry,  $x_2$ ), and that their respective cost functions are, as before,  $c_1(x_1)$  and  $c_2(x_1, x_2)$ . As a result, the equation of the production possibility locus is

$$c_1(x_1) + c_2(x_1, x_2) = R$$

For convenience let us use  $k$  as a parameter measuring the strength of the (marginal) externality.<sup>12</sup> Assume first that there are diminishing returns (increasing costs) in the production of the two outputs, and that there are no marginal external effects so that  $k=0$ . (At the margin industry I's output produces no smoke or smoke is harmless to industry II.) In that case it is easy to show that the production possibility locus must satisfy  $dx_2^2/dx_1^2 < 0$ , i.e., that the locus must assume the general shape  $AC_0B$  in Figure 2 with the concavity property required by the second-order conditions.

Now, suppose that the activity of in-

<sup>12</sup> E.g.,  $k$  may be interpreted as  $\partial^2 c_2 / \partial x_1 \partial x_2$ , i.e., the additional marginal resources cost of output 2 resulting from a unit increase in output 1.

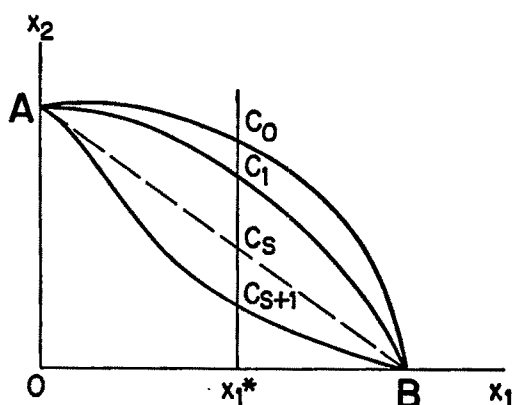


FIGURE 2

dustry I does produce some external damage ( $k > 0$ ). What happens to the production possibility locus? First I will argue that neither of its end points,  $A$  or  $B$ , will normally be affected. At point  $B$ , laundry output,  $x_2$ , is zero. Hence, no matter how much smoke is produced, there is no laundry output to be damaged. Point  $B$  is therefore invariant with the magnitude of  $k$ . Similarly, at  $A$ , the smoke creating output is zero. Consequently, no matter how smoky the process of producing output II may be (no matter how large the value of  $k$ ) the total smoke emitted will be (output  $x_1$ )  $\cdot$  (smoke per unit of output)  $= 0$ , since the first of these factors is zero. Thus the position of point  $A$  remains invariant with the magnitude of  $k$ .

The effect on intermediate points such as  $C_0$  on the locus is quite different. As  $k$  increases it takes increasing quantities of resources to produce a given volume of laundry. Thus, with any fixed value of  $x_1$ , say  $x_1^*$ , as  $k$  increases, the quantity of laundry that can be turned out with a given quantity of resources,  $R$ , must decline. Point  $C_0$  will be pushed down to some lower point,  $C_1$ . With a still greater value of  $k$  it will be lowered still further. As smoke damage increases without limit it will take larger and larger quantities of resources to turn out a given quantity of laundry and eventually we approach a

limit point  $\gamma$  on the horizontal axis, at which it is no longer possible to produce clean clothes with any finite quantity of resources.

Now draw in straight line segment  $AB$  whose position does not vary with  $k$  since neither  $A$  nor  $B$  does. It is clear that as  $k$  increases we will eventually come to some point  $C_s$  beyond which all remaining points in the sequence  $C_{s+1}$ ,  $C_{s+2}$ , . . . lie below  $AB$ . Beyond this point, obviously, the second-order conditions must be violated, as the production possibility curve approaches the axes,  $AOB$ .

Thus we see that the presence of sufficiently strong detrimental externalities will generally produce a violation of the second-order conditions. Only in the presence of insignificant externalities can one have any degree of confidence that the convexity conditions will hold.<sup>13</sup>

It is easy to offer an intuitive reason indicating how the presence of externalities increases the likelihood of a multiplicity of maxima, a reason that suggests that the problem is very real and potentially very serious in practice. Where a particular activity reduces the efficiency of another it becomes plausible that the optimal level of that activity, at least at some particular locations, is zero. If there are one hundred possible locations for the plants of a smoke-producing industry the worst possible solution might be to place some plants in each candidate location. Any solution leaving at least some combination of smoke-free areas may be preferable, and may well constitute a local maximum.

To make the point more concretely, suppose we are dealing with an island separated by a ridge of mountains that pre-

<sup>13</sup> The analysis can be extended to the case of  $n$  activities and externalities that enter utility as well as production functions. The analysis here confines itself to externalities producing inefficiencies on the production side following a suggestion of Jacob Marschak that the argument is more persuasive if framed in these terms.

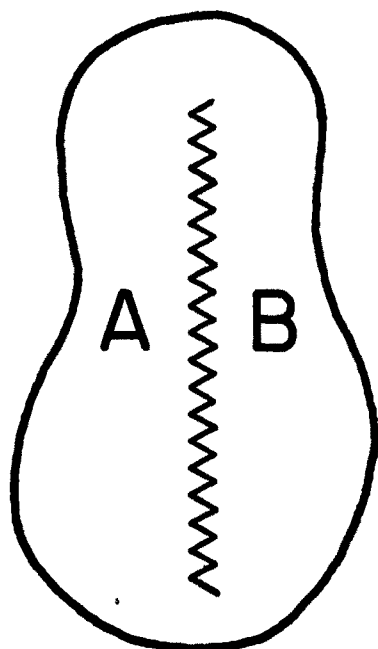


FIGURE 3

vent smoke from going from one side to the other (Figure 3). Let  $S_a$  and  $S_b$  be the volume of smoke-producing activity located on the two respective sides of the island, and let  $P_a$  and  $P_b$  be the corresponding number of residents living there. Let  $S_a + S_b = S$  and  $P_a + P_b = P$ . Then, if the social cost of the smoke is great enough, there will obviously be at least two local optima:  $(P_a = P, P_b = 0, S_a = 0, S_b = S)$  and  $(P_a = 0, P_b = P, S_a = S, S_b = 0)$ . For either of these arrangements keeps the smoke and the people apart. This does not mean, of course, that the two solutions are equally desirable. If  $A$  offers great scenic attractions while  $B$  is closer to raw materials we may expect the former of the two local maxima to be preferable. We cannot preclude the possibility of a third (interior) maximum, for once there is some industrial activity on each of the two sides of the island there may be some least cost distribution of people and industrial activity. But we see that we may well expect to encounter *at least* two local

maxima. With more separated locations and more sources of externalities the number of combinations of zero-valued variables that constitute local maxima may well grow astronomically.

The presence of a number of local maxima clearly means that an "improvement" may merely represent a move toward some minor peak in the social welfare function and it can, therefore, impose serious opportunity losses on society. All in all, we are left with little reason for confidence in the applicability of the Pigouvian approach, literally interpreted. We do not know how to calculate the required taxes and subsidies and we do not know how to approximate them by trial and error.

#### VII. An Alternative Approach—Adjustment of Taxes to Achieve Acceptable Externality Levels

There is an alternative approach to the matter that seems perfectly natural. On issues as important as those we are discussing, given the limited information at our disposal, it is perfectly reasonable to act on the basis of a set of minimum standards of acceptability. If, say, we treat the sulphur content of the atmosphere as one of the outputs of the economic system, it is not unreasonable to select some maximal level of this pollutant that is considered satisfactory and to seek to determine a tax on the offending inputs or outputs capable of achieving the chosen standard. This is precisely the approach employed in the formulation of stabilization policy, where it is decided that an employment rate exceeding  $w$  percent and a rate of inflation exceeding  $v$  percent per year are simply unacceptable, and fiscal and monetary measures are then designed accordingly.<sup>14</sup>

<sup>14</sup> As this discussion indicates, I join Welicz in refusing to abandon externalities policy entirely to Little's "administrative decisions" (p. 184) or to Ralph Turvey's "applied economist" (p. 313). For further discus-

The advantages (as well as the limitations) of this approach are clear—unlike the Pigouvian procedure, it promises to be operational because it requires far less information for its implementation. Moreover, it utilizes global measures and avoids direct controls with all of their heavy administrative costs and their distortions of consumer choice and inefficiencies. It does not use the police and the courts as the prime instrument to achieve the desired modification of the outputs of the economy. Its effects are long lasting, not depending on the vigor of an enforcement agency, which all too often proves to be highly transitory. Unlike most other measures that have been proposed in the area it need not add to the mounting financial burdens of the state and local governments. Finally, it can be shown that, unlike any system of direct controls, it promises, at least in principle, to achieve decreases in pollution or other types of damage to the environment at minimum cost to society.<sup>15</sup>

---

sion see Baumol and Wallace Oates. For an earlier proposal that is very similar in spirit, see John H. Dales, ch. 6.

<sup>15</sup> This proposition has been suggested elsewhere (see, for example, Kneese and Bower, chs. 5 and 7; Larry Ruff, p. 79), and will be fairly obvious to anyone familiar with the analysis of the allocative effects of price changes and their efficiency properties. Specifically, suppose it is desired to reduce the pollution content of a river by  $k$  percent. Obviously a  $k$  percent reduction in the number of gallons emitted by each of the plants discharging wastes into the river will generally not be the desired solution. The theorem in question then asserts the following:

*Given the production of any desired vector of final outputs by the plants along the river, a tax per gallon of effluent sufficient to reduce the overall pollution content of the river to the desired level will automatically achieve this decrease at minimum total cost to all plants combined.*

The proof of the theorem is a straightforward exercise in constrained maximization (see Baumol and Oates). It works, of course, because the lower the marginal cost of reduction in pollution outflows of a particular plant, the larger the reductions it will pay it to undertake to avoid the corresponding tax payment.

What is surprising about the proposition, if anything, is that, unlike many results in welfare analysis, it does not require the firms along the river, or any other firms,

One can expect an acceptability criterion procedure to be operational because policy makers think quite naturally in terms of minimum acceptability standards, and while it is no doubt an exaggeration to say that they can arrive at them easily, there are all sorts of precedents indicating that such standards can be decided upon in practice.

Though we are unlikely to be able to determine in advance precisely a set of tax values that will achieve the desired output standards, the output level achieved by a given tax arrangement is readily observed and, at least in principle, it is possible to learn by trial and error, continuing the direction of change of any tax modifications that turn out to bring outputs closer to their target levels. Since the procedure is a satisficing rather than a maximizing approach the possibility of a multiplicity of maxima is not relevant.

That is to say, one generally expects a considerable number of solutions to satisfy a particular set of acceptability conditions (various resource allocation patterns may be able to achieve a given set of reductions in pollution levels) *whether or not the second-order conditions are satisfied*. If several of these do so, then the essence of the satisficing approach is that one simply utilizes the first of the acceptable solutions that is discovered. One gives up any attempt to achieve any standard of optimality (other than minimization of cost<sup>16</sup> for a given degree of protection of the environment) and rests content with *any* solution that happens to satisfy the standards that have been selected.

---

to be perfect competitors, nor does it have to assume that they maximize profits rather than share of market or growth or some other target variable. All it requires is that the firms wish to produce whatever output they select at minimum cost to themselves.

<sup>16</sup> Of course it is conceivable that there may be more than one local cost minimum. In that case an effluent charge that yields an acceptable pollution level may not yield the global cost minimum. This may be something that practical policy simply has no way of avoiding.

Thus, the acceptability criterion approach does not dispose of the difficulties involved in finding a true optimum—rather it sweeps those difficulties under the rug. Even with pollution reduced to acceptable levels, there will remain the possibility that the (undiscovered) global optimum offers us a world far better than what we have managed to achieve—if only we knew how to attain it. But if we permit ourselves to be paralyzed by councils of perfection we may have still greater cause for regret.

It may be that with time we can learn to improve the workings of a set of standards of acceptability. If, say, it turns out to be unexpectedly cheap to attain the initial pollution standards, it may be reasonable to tighten the standards on the presumption that marginal costs will not yet have equalled the marginal social benefits. Successive modifications in the criteria based on experience and revaluation may produce results that on the whole are not too bad.

If firms are put on notice that the acceptability standards may well be modified in the future this may lead them to construct what George Stigler describes as more flexible plants,—plants which are designed to keep down the cost of response to changes in standards. Of course, flexibility itself is not costless. However, it may be precisely what is appropriate for a society which is only beginning to learn how to grapple with its environmental problems.

#### APPENDIX

##### *Buchanan, Stubblebine and Taxation of Both Parties to an Externality*

Buchanan and Stubblebine have raised objections to the Pigouvian solution similar to those offered by Coase (see fn. 2, above). Much of their discussion deals with the case where voluntary negotiation in the presence of externalities will lead automatically to a

Pareto optimum. As already admitted, in this case a Pigouvian tax will only cause trouble. However, the authors also appear to offer an argument against the Pigouvian tax for the case in which negotiation is absent.

Their argument, if I understand it correctly, is that after industry I adjusts to a Pigouvian tax on its output, for that industry the marginal yield of an increase in  $x_1$  is zero. However, for industry II, at the point  $\gamma$  the marginal yield of  $x_1$  is  $c_{21} < 0$ . There must, consequently, be potential gains from trade between the two industries. They state:

So long as  $[(\partial c_2/\partial x_1)/(\partial c_2/\partial x_2)]$  remains nonzero, a Pareto-relevant marginal externality remains, despite the fact that the full 'Pigouvian solution' is attained. The apparent paradox here is not difficult to explain. Since, as postulated, [II] is not incurring any cost in securing the change in [I's] behavior, and since there remains, by hypothesis, a marginal diseconomy, further 'trade' can be worked out between the two parties. . . . The important implication to be drawn is that full Pareto equilibrium can never be attained via the imposition of unilaterally imposed taxes and subsidies . . .

[Section III, pp. 382–83]

No doubt this is true—in a competitive situation two interrelated industries can generally increase their joint profits ("gain from trade") by collusion at the expense of the general public. In the case under discussion, if the output of  $x_1$  is reduced it is true that industry I will lose nothing and industry II will gain  $c_{21}$ . However, society as a whole will experience no net gain.

Since the analysis deals exclusively with resource *allocation* we must assume that the labor released by the reduced value of  $x_1$  will be employed elsewhere to produce more of some other output or more leisure. Consequently, the goods or services represented by the  $t$  units in taxes must be redistributed to the general public either by remission of another tax, increased provision of government services or some other means.

We may now evaluate the consequences of a unit increase in the output of  $x_1$  on the entire society by summing up the direct effects on each of the three groups immediately

	Industry I	Industry II	Consumers	General Public
Incremental gain or revenue	$p$		$u_1 = c_{11} + t$	$t = c_{21}$
Incremental cost	$(c_{11} + t) = (c_{11} + c_{21})$	$c_{21}$	$p$	

concerned: industry I, industry II, consumers, and the consequences of the tax receipts for the general public (which encompasses all consumers and producers, including those already mentioned). These are shown in the table above. Adding up the incremental gains and revenues we see that the net social gain is zero, precisely as optimality requires. There is only a redistribution from industry II to the general public.

In a recent letter Buchanan comments:

As for the nonoptimality of a unilaterally imposed tax, the problem here is that income effects enter to make the benefit-receiving side change behavior so that still further adjustments would be necessary . . . Our point was that this new position would not be one of full equilibrium if income effects enter. The laundries would now find that they secure the benefits of cleaner air without cost to themselves. Presumably this would make them do more laundry. This change in behavior would in turn change the apparent optimal solution. Admittedly, the imposed solution qualifies as Pareto-optimal if further trading is prohibited. And here Pareto-equilibrium does take on a different meaning from Pareto-optimal. Gains-from-trade exist, as you agree and, once these take place, we are not in an optimal solution.

In this paper I deal with the case where trading fails to take place not because it is prohibited, but because (as seems characteristic of our most important externalities problems in reality) large numbers make trading virtually impossible to arrange (where have we seen automobile drivers pay one another to cut down their exhaust?). Moreover, one must distinguish between the role of Buchanan's income effect and that of "further trading." Of course, further trading can destroy the optimality of the results achieved by a Pigouvian tax. For, as just

argued, in that case the two affected groups gain by exploiting the community. On the other hand, the "income effect"—the influx of laundries near the factory as clean air becomes cheaper is precisely the reason a tax on the smoke producer alone can lead *everyone* to behave Pareto optimally (see Section III).

## REFERENCES

- W. J. Baumol, "External Economies and Second-order Optimality Conditions," *Amer. Econ. Rev.*, June 1964, 54, 358-72.
- and W. E. Oates, "The Use of Standards and Pricing for Protection of the Environment," *Swedish J. Econ.*, Mar. 1971, 73, 42-54.
- J. M. Buchanan, "External Diseconomies, Corrective Taxes and Market Structure," *Amer. Econ. Rev.*, Mar. 1969, 59, 174-7.
- and W. C. Stubblebine, "Externality," *Economica*, Nov. 1962, 29, 371-84.
- R. H. Coase, "The Problem of Social Cost," *J. Law Econ.*, Oct. 1960, 3, 1-44.
- J. H. Dales, *Pollution, Property, and Prices*, Toronto 1968.
- O. A. Davis and A. Whinston, "Externalities, Welfare and the Theory of Games," *J. Polit. Econ.*, June 1962, 70, 241-62.
- and —, "On Externalities, Information, and the Government-Assisted Invisible Hand," *Economica*, Aug. 1966, 33, 303-18.
- F. T. Dolbear, Jr., "On the Theory of Optimum Externality," *Amer. Econ. Rev.*, Mar. 1967, 57, 90-103.
- A. V. Kneese and B. T. Bower, *Managing Water Quality: Economics, Technology, Institutions*, Baltimore 1968.
- I. M. D. Little, *A Critique of Welfare Economics*, 2d ed., New York 1957.
- H. Mohring and J. H. Boyd, "Analyzing 'Externalities': 'Direct Interaction' vs. 'Asset

- Utilization' Frameworks," *Economica*, forthcoming.
- R. A. Meyer, Jr., "Externalities as Commodities," *Amer. Econ. Rev.*, Sept. 1971, 61, 736-40.
- D. W. Pearce and G. S. Sturmev, "Private and Social Costs and Benefits: A New Terminology," *Econ. J.*, Mar. 1966, 76, 152-57.
- A. C. Pigou, *The Economics of Welfare*, 4th ed., London, 1932.
- C. R. Plott, "Externalities and Corrective Taxes," *Economica*, Feb. 1966, 33, 84-7.
- R. D. Portes, "The Search for Efficiency in the Presence of Externalities," forthcoming.
- L. E. Ruff, "The Economic Common Sense of Pollution," *Publ. Interest*, spring 1970, 19, 69-85.
- D. A. Starrett, "Fundamental Non-Convexities in the Theory of Externalities," Harvard 1971, unpublished.
- G. J. Stigler, "Production and Distribution in the Short Run," *J. Polit. Econ.*, June 1939, 47, 305-27.
- R. Turvey, "On Divergences Between Social Cost and Private Cost," *Economica*, Aug. 1963, 30, 309-13.
- S. Wellicz, "On External Economies and the Government-Assisted Invisible Hand," *Economica*, Nov. 1964, 31, 345-62.

# The Effects of Minimum Wages on the Distribution of Changes in Aggregate Employment

By MARVIN KOSTERS AND FINIS WELCH\*

Despite a generation of experience with legal minimum wages and numerous attempts to analyze their impact, there remains wide disagreement on who is affected and how they are affected.<sup>1</sup> These differences in opinion reflect the inconclusive and often inconsistent nature of much of the evidence. Studies of minimum wage effects have varied widely in terms of the kinds of effects examined, methodology, data sources, and interpretation. Some studies "find" a significant impact while others detect no important influence on employment.<sup>2</sup>

\* U.S. Cost of Living Council; Graduate Center of the City University of New York and the National Bureau of Economic Research, respectively. This research was performed at the RAND Corporation pursuant to a contract with the U.S. Office of Economic Opportunity, with some support from the National Bureau of Economic Research Faculty Fellowship to Welch. The opinions expressed herein are those of the authors and should not be construed as representing the opinions or policy of any agency of the United States Government. We are indebted to Sam Bowles, Steve Carroll, and Victor Fuchs for valuable comments, and to Thomas Moore for supplying the minimum wage data.

<sup>1</sup> For example, A. Philip Randolph states that increasing the minimum wage would do most to raise Negroes out of poverty, while Milton Friedman opines that the minimum wage law probably does Negroes the most harm of all the laws on the statute books in this country. In George Meany's view, youth unemployment is not due to provisions of the Fair Labor Standards Act, while Yale Brozen argues that increased minimum wages lead to teenage job loss, barring them from opportunities to learn job skills and impairing their work attitudes and ambitions.

<sup>2</sup> Since the primary purpose of this study is to present new empirical evidence on minimum wage effects, previous studies cannot be adequately summarized and evaluated here. Our methodology and the effects we analyze are sufficiently different to make direct comparisons of results difficult. So we have chosen to reference a number of relevant studies without making detailed comparisons.

One of the principal issues common to each of these studies is the manner in which other influences that are always at work and which tend to obscure the effects of changes in minimum wage legislation, are taken into account. For example, it is apparent from the employment and unemployment data that changes in the labor market are predominantly generated by vagaries in the pace of economic activity superimposed on a generally rising trend in productivity, incomes, and employment. As the pace of economic activity varies, effects are not distributed evenly among productive inputs. Our objective is to analyze the way changes in aggregate employment are distributed among demographic groups and to see how minimum wage legislation has altered this distribution. Historical experience leads us to expect that employment is likely to vary more than the stock of machines, and the purchases of machines (investment) may vary even more than employment. Among different classes of workers the impact of variation in economic activity is also uneven: the distribution of employment is affected by both cyclical and secular forces which are not necessarily congruent. Industrial and occupational requirements, labor force participation trends, and the age composition of the population, influence employment patterns only gradually. Changes in the level and coverage of minimum wages are abrupt, and may be only indirectly related to secular forces. If these minimum wage changes have differentially affected employment oppor-

tunities for demographic groups, it should be possible to obtain evidence of these effects, after netting out cyclical movements and taking long-term trends into account, and that is our objective.

This is an analysis of changes in the distribution of employment across age-color-sex classes. A model is developed to characterize employment patterns of different demographic groups through time. It allows for growth of aggregate employment along a fluctuating path and for long-term trends in the composition of total employment. In this framework, we estimate the effects of minimum wage legislation on the distribution of employment and on the distribution of changes in employment between whites and nonwhites, between males and females, and between teenagers and adults. We do not estimate the effects of minimum wages on total employment, instead employment is assumed to be given while we focus upon its distribution among various groups.

### I. The Model

Our model relies heavily on the distinction between persistent employment trends and deviations from trend. It is of the form:

$$(1) \quad E_{it} = \gamma_i E_{pt} + \beta_i E_{rt} + u_{it},$$

where  $E_{it}$  is the number of workers in the  $i$ th age-color-sex class employed in period  $t$ ;  $E_{pt}$  and  $E_{rt}$  are aggregate employment measures. Operationally defined,  $E_{pt}$  is an extrapolation of previous employment levels and is considered as a measure of "normal" employment; that is, it is our estimate of the level of employment consistent with the economy's long-run trend. The term  $E_{rt}$  is the deviation between actual aggregate employment,  $E_t$ , and the projection,  $E_{pt}$ .<sup>3</sup> The coefficients,  $\gamma_i$  and  $\beta_i$ ,

<sup>3</sup> The operational definition is:

$$E_{pt} = (1/2) E_{t-4} (1+r)^4 + (1/3) E_{t-8} (1+r)^8 \\ + (1/6) E_{t-12} (1+r)^{12}$$

are, respectively, the share of the  $i$ th group in normal and transitional employment. They are obviously determined by an economic system and can be affected by many things. Here they are assumed to be simple exponential functions of time to capture secular effects. The  $\gamma$  and  $\beta$  are also assumed to be dependent on the minimum wage level and the extent of coverage. The residual,  $u_{it}$ , is always with us.

The functional definition of employment shares is introduced into the model as follows:

$$(2) \quad \log \gamma_{it} = \log \gamma_{oi} + \eta_{pi} \log M_t \\ + t \log (1+r_i)$$

$$(3) \quad \log \beta_{it} = \log \beta_{oi} + \eta_{ri} \log M_t \\ + t \log (1+r_i)$$

The trend factor,  $r_i$ , is the long-term growth rate of the employment share for the  $i$ th class. The variable  $M_t$  is the "effective" minimum wage and is operationally defined in a later section, and  $\eta_{pi}$  and  $\eta_{ri}$  are elasticities of employment shares with respect to the minimum wage. The data used to estimate the model are quarterly averages of seasonally adjusted employment from 1954 through 1968 for eight age-color-sex classes.

### Employment Shares

Our hypothesis is that the composition of changes in employment differs from the normal employment distribution that would exist under stable demand. Short-term fluctuations reflect the hiring, retention, and layoff decisions of firms as they adjust to changes in product demand and

---

where  $E_{t-i}$  represents aggregate employment in the  $(t-i)$ th quarter and  $r$  is an estimate of the long-term quarterly growth rate in aggregate employment. The estimated quarterly growth rate is 0.0035 as reported in the Appendix along with a description of how it is estimated. Thus  $E_{pt}$  is a weighted average of three projections based on observations one, two, and three years in the past. The weights are admittedly arbitrary, but experiments with four alternative sets of weights showed only slight changes in resulting estimates.

there is no reason why these effects would be uniform. For example, hiring costs differ among classes of workers; it is widely held that costs of hiring low-productivity, low-wage workers are generally less than for more highly paid workers.<sup>4</sup> Further, since performance in many jobs depends on the amount of firm-specific training acquired by workers, firms often invest in this sort of training.<sup>5</sup> But, if a firm has an investment in a worker it suffers a capital loss when his employment terminates. The implication of these differences in hiring and firing costs is that relatively low-wage, low-productivity workers and those with little firm-specific investment are likely to share disproportionately in short-term changes in aggregate employment, i.e., they are more "marginal."

The employment projection,  $E_{pt}$ , can be viewed as a normal level of employment toward which labor input combinations have been brought more nearly into long-term equilibrium than would be the case for employment levels during periods of change. The set of  $\gamma$ 's represents the distribution of normal employment among age-color-sex classes, so that  $\gamma_i$  is the fraction of normal aggregate employment accounted for by the  $i$ th class. The estimated fraction is, of course, constrained by a constant trend factor and subject to a functional dependence on  $M$ .

Transitional employment,  $E_{\tau t}$ , is defined as current less normal employment and reflects shifts during a transitional period. Thus, during strong short-term expansion, transitional employment is positive and becomes negative during a slowdown or contraction. The set of  $\beta$ 's represents the distribution of transitional employment among classes of workers, and  $\beta_i$  is the

fraction of transitional employment accounted for by the  $i$ th class.

Abstracting from trends and at a given minimum wage level, the ratio of the share of a group in transitional employment to its share in normal employment is a convenient summary measure of the cyclical sensitivity of employment for a class of workers. This measure,  $\beta/\gamma$ , which we call the coefficient of marginality, is the elasticity of employment in a group with respect to deviations in aggregate employment about its normal path.<sup>6</sup>

This simple model has the capacity to produce the kinds of changes in the distribution of employment that are well known to all observers: some share much more than others in contractions and expansions. For example, if  $\beta_i/\gamma_i > 1$ , employment of workers in the  $i$ th class would vary more than proportionately with the aggregate. But if  $\beta_i/\gamma_i < 1$ , fluctuations would be dampened. In this sense the model is only descriptive and, in fact, is somewhat similar to the last-hired first-fired models.

In addition to its simplicity, the value of this model is that it can be treated as "economic" instead of being purely descriptive. Our interest is in minimum wage effects, so other variables are relegated to trend. Also, we do not in any way other than pure speculation—except for the obvious case of minimum wage legislation—address the question of why transitional, relative to normal, shares are larger for some groups than others. We would expect that workers who are marginal ( $\beta/\gamma > 1$ ) to the workforce may be characterized by little investment in firm-specific human capital and relatively low hiring costs. Conversely, workers in a class with a co-

<sup>4</sup> At this point, this is only an assertion which turns out to be consistent with our empirical observation that low-wage workers are more marginal to the workforce.

<sup>5</sup> See Gary Becker for a discussion of investment in human capital and the distinction between general and firm-specific investment.

<sup>6</sup> Assume that employment has been constant for several periods such that  $E_{pt} = E_t$  and  $E_{\tau t} = 0$ . Let  $E_{t+1} = (1 - \alpha)E_t$  to reflect a 100  $\alpha\%$  decline in employment. And let  $E_{pt+1} = E_t$  with  $E_{\tau t+1} = -\alpha E_t$ . Employment in the  $i$ th group is  $E_{it} = \gamma_i E_t$  and in the next period becomes  $E_{it+1} = \gamma_i E_t + \beta_i (-\alpha) E_t$ . The percentage change  $(E_{it+1} - E_{it})/E_{it} = (-\alpha)\beta_i/\gamma_i$ .

efficient of marginality less than unity can be viewed as intramarginal and are likely to be relatively expensive to hire and to possess relatively greater investment in firm-specific human capital.

It would be inappropriate to regard the  $\gamma$ 's and  $\beta$ 's as constant over the long term. For example, rising labor force participation rates for women lead to an increase in their share of aggregate employment. Similarly, an increase in the teenage fraction of the population increases teenage shares. The relative growth rates,  $r_i$ , allow for the influence of these and other long-term factors.<sup>7</sup>

### *Minimum Wages*

The way changes in legislated minimum wages alter the distribution of changes in aggregate employment is an empirical question. But speculation can suggest the nature of these effects.

The effects upon shares of normal employment are most obvious. To a competitive, profit-maximizing firm, a minimum-wage law precludes hiring workers whose productivity is less than the minimum wage. As a first-order approximation, the percentage change in employment of a particular group of workers is given by the product of the demand elasticity for the group's services and the proportionate increase implied by the minimum wage over the wage that otherwise would have existed. Barring large discrepancies across groups of workers in demand elasticities, we would simply expect unemployment effects to be greatest for those who otherwise would have earned the lowest wages.

There is one added dimension, since our analytical unit is a demographic group for which there is no presumption of intra-group homogeneity. We should consider wage distributions instead of single values. Nonetheless, we expect the unemployment

effects to be greatest for groups with low average productivity and that, as a result of minimum-wage legislation, their employment shares should fall with compensating increases in the shares of those who would otherwise earn more.

The effect of minimum wages upon the distribution of changes in aggregate employment is less obvious than for normal employment. Consider a business cycle as a sequence of rising then falling demand functions for each of the demographic groups. Factors determining the extent of these shifts include hiring-firing costs and skill specificity as discussed above. As demand falls for a particular group, the extent of the associated change in employment is determined by two factors: the extent of the demand change and the downward flexibility of wages. But, nothing is less flexible downward than a legislated wage. If the minimum wage is an effective downward constraint—as would be true of demographic groups with low average productivity—then it serves to increase employment fluctuations associated with demand changes.

Our hypothesis is that *minimum wages serve to reduce shares of normal employment and to increase shares of transitional employment for those groups of workers with low average productivity.*

The minimum wage variable used here,  $M_t$ , is defined as the legal minimum relative to the average wage in manufacturing, multiplied by the percent of total employment covered by the legislation. Deflation by manufacturing wages reflects the erosion of the real minimum wage between legislative adjustments that results from increases in workers' productivity and the general price level. The coefficients  $\eta_{pi}$  and  $\eta_{ti}$  in equations (2) and (3) are, respectively, elasticities of the share of normal and transitional employment with respect to the effective minimum wage. Thus,  $\eta_{ti} - \eta_{pi}$  is the elasticity of the coefficient of

<sup>7</sup> Estimated relative growth rates are reported in the Appendix.

marginality. For those groups with low productivity relative to the average, our hypothesis is that  $\eta_r - \eta_p > 0$  so that increases in the minimum wage increase their marginality.

While no other short-term factors influencing employment shares are explicitly analyzed, their combined influence is, of course, reflected in the employment share estimates. Excluding them from the model need not obscure the influence of the minimum wage unless their influence is highly irregular across cycles or correlated with minimum wage changes.

## II. The Estimates

Once employment shares are allowed to be functionally dependent upon the minimum wage, the model is non-linear and contains four parameters for each equation which represents a demographic group. There are eight groups corresponding to the intersection of the three binary classifications: age (teenage-adult); color (white-nonwhite); and sex.<sup>8</sup> The employment data refer to quarterly averages for the period 1954-68. The estimation technique is non-linear least squares and is described in the Appendix. Here, only summary statistics are reported.

### *Employment Share Estimates*

Estimates of employment shares ( $\gamma_{it}$  and  $\beta_{it}$ ) will vary over time because of the trend and because of changes in the effective minimum wage. Averages for the

<sup>8</sup> Since the eight age-color-sex classes together exhaust aggregate employment, one of the equations is redundant. The consistency question could be resolved in several ways: 1) One of the equations could be discarded, 2) the entire set of equations could be estimated subject to a consistency constraint, or 3) the consistency constraints could be used to normalize the estimates. Because the specification is non-linear in the minimum wage variable and there was no obvious way to impose the constraints in the estimating procedure, we have normalized the average employment share estimates (Table 1) and reported deviations from other implicit constraints (see the Appendix).

TABLE 1—AVERAGE SHARES OF NORMAL AND TRANSITIONAL EMPLOYMENT AND AVERAGE COEFFICIENTS OF MARGINALITY: QUARTERLY AVERAGE OF U.S. CIVILIAN EMPLOYMENT 1954-68

Employee Group	Normal Employment, $\gamma$	Transitional Employment, $\beta^a$	Coefficient Marginality $\bar{\beta}/\gamma^a$
Adults (20+)	.934	.779	0.83
White Males	.570	.425	0.75
Nonwhite Males	.058	.079	1.34
White Females	.269	.240	0.89
Nonwhite Females	.040	.036	0.90
Teenagers (16-19)	.063	.221	3.51
White Males	.032	.117	3.66
Nonwhite Males	.004	.018	4.56
White Females	.025	.074	2.95
Nonwhite Females	.002	.011	4.97

Source: See note to Table 2.

<sup>a</sup> Unconstrained estimates of  $\bar{\beta}$  sum to 1.056. The estimates shown are normalized by dividing the unconstrained estimates by 1.056.

1954-68 interval are reported in Table 1. The overwhelming feature of these estimates is the relative vulnerability of teenagers to employment change. A 1 percent decline in aggregate employment is accompanied by a 3.5 percent decline in teenage employment compared to a 0.8 percent decline for adults. Teenagers account for 6.3 percent of normal and 22.1 percent of transitional employment.

These estimates of relative employment stability imply that during a contraction, teenagers are four times as likely to lose their jobs as adults. That fluctuations in teenage employment are more pronounced than for adults is consistent with the prediction of the theory of firm-specific human capital; teenagers are typically less skilled and have less work experience. Within the teenage class, nonwhites appear to be most vulnerable to changes in aggregate employment, the estimate being that a 1 percent decline in aggregate employment reduces nonwhite teenage employment by 5 percent.

The pattern of employment change among adults is also of interest. Apparently both white and nonwhite female adults experience employment variations that are proportional to the aggregate. For males, on the other hand, white employment appears to be considerably more stable than for nonwhites. In fact, our estimate is that during a contraction in aggregate employment a nonwhite adult male is 1.8 times as likely to lose his job as a white adult male. The converse is also true, a point that helps to explain the remarkable rise in median nonwhite family income (relative to white) during the last decade and underscores the vulnerability of this gain to a slackened pace of economic activity.<sup>9</sup>

#### *Minimum Wage Estimates*

Since teenagers are marginal to the workforce, they would be expected to fare especially well during expansionary periods such as the decade from 1958 to 1968. Yet this is not reflected in their unemployment rates.<sup>10</sup> If changes in minimum wage legislation during this period have had a more pronounced influence on employment of teenagers than adults, these changes may be capable of explaining a large part of the divergent trends in adult and teenage employment patterns. Between 1954 and 1968, legal minimum wages increased from

\$0.75 to \$1.60 per hour, and relative to average hourly earnings in manufacturing they increased from .42 to .53. In addition, expansion of coverage over that period led to 62 percent of aggregate employment covered by minimum wage legislation in 1968 compared to 46 percent in 1954.

It is possible that part of the sustained high level of teenage unemployment over this period is attributable to the postwar baby boom that resulted in a large relative increase in the teenage population.<sup>11</sup> Because the increase in the teenage workforce was large, it may have been difficult for the economy to provide sufficient teenage employment opportunities to absorb the increase. The long-term trend adjustments in our model are based on trends in employment that occurred during the period, however, and the estimates of minimum wage effects obtained from the model pertain to changes in employment patterns among those employed. Rising minimum wages, in addition, may have led to lower levels of aggregate employment than would otherwise have occurred and resulted in crowding of low-productivity workers into sectors not covered by minimum wage legislation, but these effects are not explored in this analysis.

Our estimates of minimum wage elasticities are reported in Table 2. Recall that these estimates do not test for employment effects *per se*, rather, they take total employment as given and test to see if the level of the effective minimum wage affects the coefficients of marginality for each of the groups. Interestingly enough, the estimated pattern of the impact of minimum wage increases on the incidence of employment fluctuations is quite similar to the pattern of marginality reported in Table 1, so that minimum wages appear to exacer-

<sup>9</sup> Between 1958 and 1968 the relative income position of nonwhite families increased from .51 to .63 (see *Current Population Reports*). In the post-World War II period, the ratio of nonwhite to white median income has moved inversely with aggregate unemployment rates, reaching its post-1950 low in 1958, the year with the highest unemployment rate since World War II when 6.8 percent of the labor force was unemployed, and relative nonwhite family income was higher than any time before in 1968 when unemployment was 3.6 percent.

<sup>10</sup> For example, between 1958 and 1968 the unemployment rate for white adult males dropped from 5.5 to 2.0 percent and for nonwhite adult males from 12.7 to 3.9. Yet white teenage male unemployment only dropped to 10.1 from 15.7 and nonwhite teenage males dropped from 26.8 to 22.1.

<sup>11</sup> Between 1958 and 1968, for example, the teenage population increased from 9.5 million to 14.2 million or by 49.5 percent. And during the same period the adult population increased by only 13.6 percent.

TABLE 2—ESTIMATED ELASTICITIES OF EMPLOYMENT SHARES AND OF MARGINALITY COEFFICIENTS WITH RESPECT TO THE EFFECTIVE MINIMUM WAGE

Employee Group	Normal Employment $\eta_p^a$	Transitional Employment $\eta_r^a$	Coefficient of Marginality $\eta_r - \eta_p$	$R^2$
Adults				
White Males	.032	1.44	-1.47	.947
Nonwhite Males	-.004 <sup>b</sup>	-.47 <sup>b</sup>	-.47	.979
White Females	.032	-1.00 <sup>b</sup>	-1.03	.994
Nonwhite Females	-.017 <sup>b</sup>	.08 <sup>b</sup>	.10	.990
Teenagers				
White Males	-.331	2.48	2.81	.949
Nonwhite Males	-.356	3.88	4.24	.712
White Females	-.241	3.30	3.54	.925
Nonwhite Females	-.301	5.31	5.61	.744

<sup>a</sup> Since this is a "distributional theory," we have the implicit constraints  $\sum_i \gamma_i \eta_{pi} = \sum_i \beta_i \eta_{ri} = 0$ . These constraints are not imposed. The estimates are that  $\sum_i \gamma_i \eta_{pi} = .007$  and  $\sum_i \beta_i \eta_{ri} = -.233$ .

<sup>b</sup> Heuristic  $t$ -statistic less than 2.0. The  $t$ -statistics are calculated from observed changes in residual sums-of-squares as the square root of the ordinary  $F$ -statistic for testing the significance of a constraint that the coefficient on the variable in question is zero. Standard errors are not computed for the difference,  $\eta_r - \eta_p$ .

bate the existing structure. Minimum wage increases have heightened the employment vulnerability of teenagers, apparently by reducing their share of normal employment and increasing their share of transitional employment. Thus, the effect seems to be largely one of shifting teenagers from the normal to the transitional work force. Again, nonwhites are more affected than whites. For adults, the overall effect of increased minimum wages is to stabilize their employment. This increased stability occurs through shifting the incidence of employment fluctuations to teenagers, and the stabilizing effect is strongest for white adult males. It also appears that increased minimum wages may have stabilized employment for nonwhite adult males, although the effect is apparently very small and the statistical insignificance of the elasticities precludes any strong statements.

The estimates indicate that increases in the effective minimum wage have had a significant influence on the short-term em-

ployment stability of different classes of workers. But it is also important to consider how much employment stability has been affected by minimum wage changes in the range that we have experienced. Suppose, for example, that the effective minimum wage were suddenly changed from its average level in the first five-year period of our sample (1954-58) to the level prevailing in the third five-year period (1964-68). On the basis of the elasticities reported in Table 2, by how much would we predict coefficients of marginality to change from each of the groups? The predicted changes are: *white adult males, down 33 percent. Nonwhite adult males, down 12 percent. White adult females, down 25 percent. Nonwhite adult females, no change. And all teenage coefficients of marginality would more than double!*

### III. Summary and Conclusions

That employment of nonwhite workers is more sensitive to the pace of economic activity than for whites, and teenagers

more sensitive than adults, are not surprising results and these phenomena have been noted in earlier studies. What is noteworthy about our results, however, is the manner in which increases in minimum wages appear to have altered the distribution of employment and sensitivity to short-term changes in employment. By focusing on the distribution of employment and how that distribution changes over the cycle, we have developed estimates of some aspects of minimum wage effects that have not previously been analyzed.

Our evidence indicates that increases in the effective minimum wage over the period 1954-68 have had a significant impact on employment patterns. Minimum wage legislation has had the effect of decreasing the share of normal employment and increasing vulnerability to cyclical changes in employment for the group most "marginal" to the work force—teenagers. Thus, as a result of increased minimum wages, teenagers are able to obtain fewer jobs during periods of normal employment growth and their jobs are less secure in the face of short-term employment changes.

Minimum wage legislation has undoubtedly resulted in higher wages for some of the relatively low productivity workers who were able to obtain employment than these workers would have received in its absence.<sup>12</sup> The cost in terms of lost employment opportunities and cyclical vulnerability of jobs, however, has apparently been borne most heavily by teenagers. And a disproportionate share of these unfavorable employment effects appears to have accrued to nonwhite teenagers. The primary beneficiaries of the shifts in the

pattern of employment shares occasioned by minimum wage increases were adults, and among adults, particularly white adult males.

#### APPENDIX

The model is estimated one equation at a time without imposing the constraints of internal consistency,  $\sum_i \beta_{it} = \sum_i \gamma_{it} = 1$  and  $\sum_i \eta_{pi} \gamma_{it} = \sum_i \eta_{ri} \beta_{it} = 0$ . The estimation technique is non-linear. We simply iterate over values of  $\eta_p$  and  $\eta_r$  to minimize the residual sum of squares. The typical equation (omitting subscript,  $i$ ) is

$$E_t = \gamma_0 Z_{1t} + \beta_0 Z_{2t} + U_t$$

where  $Z_1$  and  $Z_2$  are constructs that are respectively proportional to  $E_{pt}$  and  $E_{rt}$ , where the factors of proportionality are of the form

$$\log(Z_{1t}/E_{pt}) = \eta_p \log M_t + t \log(1+r)$$

$$\log(Z_{2t}/E_{rt}) = \eta_r \log M_t + t \log(1+r)$$

There are 60 observations and 4 parameters,  $\gamma_0, \beta_0, \eta_{pi}, \eta_{ri}$  are estimated. For particular values of  $\eta_p$  and  $\eta_r$ ,  $Z_1$  and  $Z_2$  are computed and the OLS regression of  $E_{it}$  on  $Z_1$  and  $Z_2$  is then calculated. Iterating over  $\eta_p$  and  $\eta_r$ , that solution is selected to minimize the residual sum of squares.

The standard errors of  $\hat{\gamma}_0$  and  $\hat{\beta}_0$  are computed as though the regression of  $E_i$  on  $Z_1$  and  $Z_2$  were of the standard form except that degrees of freedom are 56 instead of 58. The standard errors of  $\hat{\eta}_{pi}$  and  $\hat{\eta}_{ri}$  are computed heuristically to measure the sensitivity of the residual sum of squares to the constraints  $\eta_{pi}=0$  and  $\eta_{ri}=0$ . Specifically, to compute the standard error of  $\hat{\eta}_p$  we first estimate the equation as described above and then impose the constraint,  $\eta_{pi}=0$  and iterate on  $\eta_{ri}$  to minimize the residual sum of squares. Let  $Q_0$  represent the unconstrained residual sum of squares and let  $Q_1$  represent the constrained sum of squares. Then compute

$$"F_{1,56}" = \frac{Q_1 - Q_0}{Q_0/56},$$

$$"t_{56}" = ("F_{1,56}")^{1/2},$$

<sup>12</sup> The misallocation of employment that results from partial coverage of minimum wage legislation, however, suggests that some workers may receive lower wages than they would have in the absence of minimum wage legislation.

and

$$\sigma(\hat{\eta}_p) = \frac{\hat{\eta}_p}{\sqrt{t_{56}}}$$

The standard error of  $\hat{\eta}_r$  is computed the same way.

The employment data are quarterly averages computed from seasonally adjusted monthly data reported in *Employment and Earnings*. The minimum wage and coverage data were provided by the U.S. Dept. of Labor and supplied to us by Thomas G. Moore. The minimum wage data are reported in Kusters and Welch.

The quarterly growth rates in employment for each of the groups are computed as observed growth rates between the year 1956-III to 1957-II and one decade later, 1966-III to 1967-II. This period was chosen because overall unemployment rates were similar in both years. The relative growth rate,  $r_i$ , is the quarterly growth rate of employment in this group relative to that of aggregate employment over the decade. They are:

	White Males	Nonwhite Males	White Females	Nonwhite Females
Adults	-.0023	-.0003	.0023	.0027
Teenagers	.0077	.0029	.0071	.0063

Quarterly growth rate in Aggregate Employment = .0035.

#### REFERENCES

- P. S. Barth, "The Minimum Wage and Teenage Unemployment," *Ind. Relat. Res. Ass. Proc., Twenty-Second Annual Meeting*, 1969, 296-310.
- G. S. Becker, *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*, New York 1964.
- Y. Brozen, "The Effect of Statutory Minimum Wage Increases on Teenage Unemployment," *J. Law Econ.*, Apr. 1969, 12, 109-22.
- , "Minimum Wage Rates and Household Workers," *J. Law Econ.*, Oct. 1962, 5, 103-09.
- and M. Friedman, "The Minimum Wage—Who Really Pays? An Interview," The Free Society Ass., Inc., Washington, D.C., Apr. 1966.
- A. F. Burns, *The Management of Prosperity*, Pittsburg 1966.
- J. E. Easley and R. M. Fern, "Minimum Wages and Unemployment of Teenagers," unpublished manuscript, North Carolina State Univ. 1969.
- L. A. Ferman, J. L. Kornbluh, and J. A. Miller, eds., *Negroes and Jobs*, with Foreword by A. P. Randolph, Ann Arbor 1968.
- H. Falk, "The Problem of Youth Unemployment," in *The Transition from School to Work*, Ind. Rel. Sec., res. report 111, Princeton 1968.
- E. Kalachek, *The Youth Labor Market*, Ann Arbor 1969.
- , "Determinants of Teenage Employment," *J. Hum. Resources*, winter 1969, 4, 3-21.
- A. Katz, "Teenage Employment Effects of State Minimum Wage Laws," *J. Hum. Resources*, forthcoming.
- D. E. Kaun, "Economics of the Minimum Wage: The Effects of the Fair Labor Standards Act, 1945-60," unpublished doctoral dissertation, Stanford Univ., 1963.
- M. Kusters and F. Welch, *The Effects of Minimum Wages on the Distribution of Changes in Aggregate Employment*, RM-6273-OEO, The RAND Corp., Santa Monica, Sept. 1970.
- G. Meany, "Appendix to Statement before the General Labor Sub-Committee of the House Education and Labor Committee on Legislation to Amend the Fair Labor Standards Act," July 29, 1970.
- T. G. Moore, "The Effect of Minimum Wages on Employment," *J. Polit. Econ.*, July/Aug. 1971, 79, 897-902.
- J. Peterson, and C. T. Steward, Jr., *Employment Effects of Minimum Wage Rates*, Washington 1969.
- G. W. Scully, "The Impact of Minimum Wages on Unemployment Rate of Minority Group Labor," unpublished paper, Ohio Univ., undated.
- L. C. Thurow, "The Determinants of the Occupational Distribution of Negroes," in

- G. Somers, ed., *Education and Training of Disadvantage Minorities*, Madison 1969.
- U.S. Department of Commerce, *Current Population Reports*, "Consumer Income," Series P-60, No. 66, Washington, Dec. 1969.
- Bureau of Labor Statistics, *Employment and Earnings and Monthly Report on the Labor Force*, Vol. 15, No. 8, Washington, Feb. 1969.
- , *Youth Unemployment and Minimum Wages*, Bull. 1657, Washington 1970.
- U.S. Department of Labor, Office of the Secretary, Report Submitted to Congress in Accordance with the Requirements of Section 4(d) of the Fair Labor Standards Act, Washington, D.C. 1963 through Jan. 1970.

# Implications of the Theory of Rationing for Consumer Choice Under Uncertainty

By PETER A. DIAMOND AND MENAHEM YAARI\*

A natural outgrowth of World War II was the development of the theory of consumer choice under several simultaneous budget constraints, representing the household's income constraint and as many rationing-point systems as might be imposed. Equally naturally, this theory has received far less attention in more recent years. However, in formulating one of the standard problems of consumer choice under uncertainty in a contingent commodity setting, we realized that this problem is mathematically equivalent to that of choice under rationing. In place of the constraints for the different point budgets, we have the equations relating consumption in the different states of nature to income in those states. Analogous to the prices of a given commodity in terms of money and the various point systems are the yields of a given security in the various states of nature. It is the purpose of the present essay to try to exploit this equivalence.

In Section I, we shall state a consumer's saving-and-portfolio-selection problem and exhibit its equivalence to the problem of consumer choice under rationing. Section II will be devoted to a restatement of the principal results of rationing theory. In Section III, these results of rationing

theory will be applied to the saving-and-portfolio problem of Section I. Finally, in Section IV, we shall consider the savings problem in the presence of a single security, for which some further results, which depend solely on the convexity of preferences, may be derived.

## I. A Saving- and Portfolio-Problem for a Consumer Facing Uncertainty

We shall consider a two-period, one-decision, consumer-allocation problem. The decision takes place in the first period and it entails the determination of the current (first-period) consumption level and the allocation of savings among various investment opportunities. Investments bear fruit in the second period at rates which are uncertain at the time of decision. Also uncertain at the time of decision is noninvestment income in the second period. We shall use the "states-of-nature" approach to formalize the way in which uncertainty enters. That is, for any given investment decision, the consumer's total income in the second period can take on one of  $n$  possible values, according to which one of  $n$  possible "states" obtains. Since we are dealing in a two-period problem, second-period consumption is equal to total income in the second period.

Let  $c_0$  be the consumption level in the first period, and let  $c_i$ , for  $i=1, \dots, n$ , be the consumption level in the second period if state  $i$  occurs. A "commodity bundle" in this framework is an  $(n+1)$ -tuple of nonnegative real numbers, of the form  $(c_0, c_1, \dots, c_n)$ . A utility  $u$  is defined on the set of all such bundles to the real

\* The authors are at Massachusetts Institute of Technology and Hebrew University, Jerusalem, respectively. The research reported here has been supported by the National Science Foundation, by the Maurice Falk Institute for Economic Research in Israel, and by the Mathematical Social Science Board. Part of the work on this paper was done while Diamond was at the Hebrew University, Jerusalem, and part while both authors were at the University of California at Berkeley. We are grateful to both institutions for their hospitality, and to the Institute of Business and Economic Research at Berkeley for the clerical services it provided.

numbers, and it is assumed to be twice continuously differentiable, increasing, and strictly quasiconcave. The utility  $u$  reflects both a possible dependence of preferences upon the state of nature and the consumer's beliefs on the likelihood of each state's occurrence. The consumer is viewed as maximizing  $u(c_0, c_1, \dots, c_n)$ , subject to the budget restrictions.<sup>1</sup>

Let  $y_0$  be the consumer's first-period wealth, and let  $y_i$ , for  $i=1, \dots, n$ , be second-period noninvestment income, if state  $i$  occurs. Assume, also, that there exist  $m$  securities and that all investments take the form of either purchases or short sales of these securities. For  $i=1, \dots, n$  and  $j=1, \dots, m$ , let  $\alpha_{ij}$  be the gross rate of return on security  $j$  in state  $i$ . Finally, let  $s_j$  be the amount (in terms of initial wealth) invested by the consumer in security  $j$ . A positive  $s_j$  indicates a purchase, and a negative  $s_j$  indicates a short sale.

The consumer's optimization problem may now be stated as follows:

$$(1) \quad \text{maximize } u(c_0, c_1, \dots, c_n)$$

subject to:

$$y_0 = c_0 + s_1 + \dots + s_m$$

$$c_1 = y_1 + \sum_{j=1}^m \alpha_{1j}s_j$$

$$\vdots$$

$$c_n = y_n + \sum_{j=1}^m \alpha_{nj}s_j$$

Using the first constraint, one may eliminate  $c_0$ . This leads to:

$$(2) \quad \text{maximize } u\left(y_0 - \sum_{j=1}^m s_j, c_1, \dots, c_n\right)$$

<sup>1</sup> Naturally the expected utility approach is a special case of this formulation where  $u$  takes the form  $\sum_{i=1}^n \pi_i u(c_0, c_i)$  where the  $\pi_i$  are probabilities. For a discussion of the expected utility approach and its relationship to the mean-variance approach, see the recent exchange among Karl Borch, Martin Feldstein, and James Tobin (1969).

subject to:

$$c_i = y_i + \sum_{j=1}^m \alpha_{ij}s_j, \quad i = 1, \dots, n$$

An example with two states of nature is shown in Figure 1. If there are two securities the consumption possibilities are shown by the plane, and the optimal consumption point by the tangency between the budget plane and the indifference surface which is shown in the figure. If there were only one security, the consumption possibilities would be represented by a line. The optimal consumption point is then the tangency between the budget line and the indifference surface. Two such possible budget lines are shown in Figure 1, both of which happen to give rise to the same optimal consumption point.

Let us now define a vector  $x$ , in the

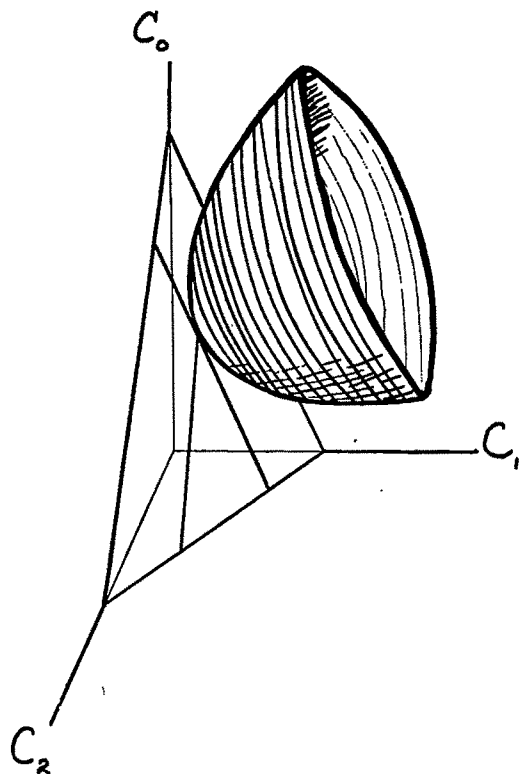


FIGURE 1

euclidean space of dimension  $m+n$ , in the following manner:

$$(3) \quad x = \begin{bmatrix} c_1 \\ \vdots \\ c_n \\ -s_1 \\ \vdots \\ -s_m \end{bmatrix}$$

Also, let  $A$  be the real  $n \times m$  matrix whose  $ij$ th entry is  $\alpha_{ij}$ , and let  $P$  be the matrix defined by:

$$(4) \quad P = [I, A],$$

where  $I$  is the  $n \times n$  identity matrix. Finally, let  $y$  be the vector, in euclidean space of dimension  $n$ , whose components are the second-period noninvestment incomes,  $y_1, \dots, y_n$ . With this notation, we may rewrite the consumer's problem as:

$$(5) \quad \text{maximize } u(x)$$

subject to:  $Px = y$ ,

where  $u$  is a utility whose relationship to  $u$  is given in (2), above.<sup>2</sup>

We now have the consumer's optimization problem stated in a form that is identical to the optimization problem of a consumer choosing a commodity bundle in a world of rationing. Thus, one is in a position to apply known results from the theory of rationing to the theory of saving-and-portfolio selection in a world of uncertainty. Before doing so, however, we ought to present a summary of the theory of rationing, because twenty years or more have elapsed since rationing theory actively occupied the thoughts of economists.

## II. Choice with Rationing

In this section we present a fairly com-

<sup>2</sup> By examining the equation  $u = u$ , one can verify that  $u$  has the usual regularity properties of a utility, namely, differentiability, monotonicity, and quasiconcavity.

plete outline of the theory of demand under rationing. Most of the results may be found in classical references (such as Paul Samuelson and Tobin (1952)), but the approach here will be somewhat different.

Let a consumer be described by a consumption set  $X$ , by a utility function  $u$ , and by an  $m$ -tuple of incomes  $y$ . The set  $X$  is assumed, for simplicity, to be a translate of the nonnegative orthant in euclidean  $n$ -space. The utility  $u$  is defined on  $X$  to the real numbers and is assumed to be continuous, increasing (in each coordinate), and strictly quasiconcave. As for the vector  $y$ , its  $i$ th component, call it  $y_i$ , represents the consumer's income in the  $i$ th budget constraint.

The environment in which the consumer operates is described by a real matrix  $P$ , whose  $ij$ th entry, call it  $p_{ij}$ , represents the price of the  $j$ th commodity in the  $i$ th budget constraint.  $P$  is a matrix of order  $m \times n$ . Trading the various incomes for one another is assumed impossible. Thus, we are led to the following choice problem for the consumer:

$$(6) \quad \text{maximize } u(x)$$

subject to:  $Px \leq y, x \in X$

Throughout the analysis, we shall assume that the maximum occurs in the interior of the consumption set  $X$ ; that is, we shall ignore the constraint  $x \in X$  in (6). Furthermore, we now wish to convert the inequality  $Px \leq y$  into an equation,  $Px = y$ . The analysis will be entirely local, in the neighborhood of the optimum. Does this fact permit us to say that all the constraints in  $Px \leq y$  that are nonbinding at the optimum have been eliminated and only the binding ones retained? To give a positive answer to this question, we must make a "nondegeneracy" assumption to the effect that, at the optimum, the shadow prices of all the constraints that hold with equality are positive.

Having made all these assumptions, we may now describe the consumer's problem as choosing  $x$  to

$$(7) \quad \text{maximize } u(x)$$

$$\text{subject to:} \quad Px = y$$

Let the optimal commodity bundle be denoted  $x(P, y)$ , and define a function  $v$  by  $v(P, y) = u(x[P, y])$ . The function  $v$  is known as the indirect utility function.<sup>3</sup> It may easily be verified that, just as in standard consumer theory, the function  $v$  has the following properties: it is continuous, increasing in  $y$ , decreasing in  $P$ , quasiconvex,<sup>4</sup> and homogeneous of degree zero. A standard revealed preference argument leads to the conclusion that the following duality exists: if  $x$  is a commodity bundle (in the given neighborhood), then the pair  $(P, y)$  at which  $x$  is chosen is the solution of the problem of selecting  $P$  and  $y$  to<sup>5</sup>

$$(8) \quad \text{minimize } v(P, y)$$

$$\text{subject to:} \quad Px = y$$

This problem may be solved by substituting from the constraint into  $v$ . In other words, let a function  $w_x$  be defined by  $w_x(P) = v(P, Px)$ . Then, minimizing  $w_x(P)$  leads to the price matrix at which  $x$  is optimal. The matrix is not unique, of course, since the prices and income of any budget can be multiplied by a positive scalar without any effect. The function  $w_x$  may be referred to as the *compensated* indirect

utility function at  $x$ , since the income levels are adjusted to permit purchase of the same commodity bundle  $x$ .

Assuming that we have the differentiability of all quantities and the ability to use the calculus in analyzing the response to changes in the parameters, let us introduce the following notation:

$$(9) \quad \begin{aligned} \partial v(P, y) / \partial p_i^j &= v_i^j & i &= 1, \dots, m \\ & & j &= 1, \dots, n \\ \partial v(P, y) / \partial y_i &= v_i & i &= 1, \dots, m \end{aligned}$$

The term  $v_i$  is, of course, equal to the Lagrange multiplier associated with the  $i$ th budget constraint in the primal maximization problem. We are assuming that  $v_i > 0$  for all  $i$ .

From the nature of the dual problems, we have a solution to (8) and, thus, where prices are positive, the first-order conditions will be satisfied. Thus, we have

$$\partial w_x(P) / \partial p_i^j = \partial v(P, Px) / \partial p_i^j = 0$$

for all  $i$  and  $j$ ; that is,

$$(10) \quad v_i^j + v_i x^j = 0 \quad \text{for all } i \text{ and } j$$

These equations imply that the quantity  $v_i^j / v_i$  is independent of  $i$ . In the standard consumer problem, this equation also holds and is a direct way of calculating the demand curves from the indirect utility function.<sup>6</sup>

The dual problem, which sees the consumer as minimizing the utility cost of attaining a given commodity bundle, is quite useful in obtaining information about the effects of compensated price and income changes on the consumer's behavior. Consider a change in  $p_k^l$ , the price of the  $l$ th commodity in the  $k$ th constraint, compensated by changing  $y_k$ , income in the  $k$ th constraint, in such a way that utility remains constant. Denote the rate of change in  $x^j$ , the demand for the  $j$ th commodity,

<sup>3</sup> Note that  $v$  is well defined only locally, i.e., for values of  $P$  and  $y$  that lead to an optimum in a given neighborhood. The entire discussion is subject to this qualification.

<sup>4</sup> The function  $v$  is quasiconvex and also strictly quasiconvex for price changes that lead to changes in demand.

<sup>5</sup> This parallels duality with a single budget constraint. The optimality of  $x$  represents the highest utility that can be obtained on the budget set. At  $x$  we also have the lowest utility among those points which are optimal for the different budgets which could be drawn through  $x$ . Obviously, no budget through  $x$  could lead to an optimum with a lower utility than that at  $x$ .

<sup>6</sup> This was pointed out by R. Roy.

as a result of this compensated change in  $p_k^l$ , by the symbol

$$\left. \frac{\partial x^j}{\partial p_k^l} \right|_{y_k-\text{comp}}$$

As in the case of standard consumer theory, this compensated rate of change is given by:

$$\begin{aligned} (11) \quad \left. \frac{\partial x^j}{\partial p_k^l} \right|_{y_k-\text{comp}} &= \left. \frac{\partial x^j}{\partial p_k^l} + \frac{\partial y_k}{\partial p_k^l} \right|_{v \text{ const}} \cdot \frac{\partial x^j}{\partial y_k} \\ &= \frac{\partial x^j}{\partial p_k^l} - \frac{v_k^j}{v_k} \frac{\partial x^j}{\partial y_k} \\ &= \frac{\partial x^j}{\partial p_k^l} + x^l \frac{\partial x^j}{\partial y_k} \end{aligned}$$

where the first equality is the definition of the compensated derivative, the second uses the constraint  $dv=0$ , and the third makes use of (10).

A direct method for obtaining the standard results for the Slutsky matrix is to show its proportionality to the matrix of second derivatives of  $w_x$ . Let us calculate these derivatives:<sup>7</sup>

$$\begin{aligned} (12) \quad \frac{\partial w_x}{\partial p_i^j} &= v_i^j + x^j v_i, \\ \frac{\partial^2 w_x}{\partial p_k^l \partial p_i^j} &= \frac{\partial v_i^j}{\partial p_k^l} + x^j \frac{\partial v_i}{\partial p_k^l} \\ &\quad + x^l \left( \frac{\partial v_i^j}{\partial y_k} + x^j \frac{\partial v_i}{\partial y_k} \right) \end{aligned}$$

Solving (10) for  $x^j$  and differentiating the resulting equation with respect to  $p_k^l$  and  $y_k$  gives:

$$\begin{aligned} (13) \quad \frac{\partial x^j}{\partial p_k^l} &= - (v_i)^{-1} \left( \frac{\partial v_i^j}{\partial p_k^l} - \frac{v_i^j}{v_i} \frac{\partial v_i}{\partial p_k^l} \right) \\ &= - (v_i)^{-1} \left( \frac{\partial v_i^j}{\partial p_k^l} + x^j \frac{\partial v_i}{\partial p_k^l} \right), \end{aligned}$$

<sup>7</sup> Note that in differentiating the first-order conditions (10),  $x^j$  is a function of  $P$  and  $y$ , whereas in differentiating  $w_x$ ,  $x$  is treated as a parameter.

$$\begin{aligned} \frac{\partial x^j}{\partial y_k} &= - (v_i)^{-1} \left( \frac{\partial v_i^j}{\partial y_k} - \frac{v_i^j}{v_i} \frac{\partial v_i}{\partial y_k} \right) \\ &= - (v_i)^{-1} \left( \frac{\partial v_i^j}{\partial y_k} + x^j \frac{\partial v_i}{\partial y_k} \right) \end{aligned}$$

Comparing (11), (12), and (13), we see that:

$$(14) \quad \left. \frac{\partial x^j}{\partial p_k^l} \right|_{y_k-\text{comp}} = - (v_i)^{-1} \frac{\partial^2 w_x}{\partial p_i^j \partial p_k^l}$$

which holds for all  $i=1, \dots, m$ . Now, the matrix of second derivatives of  $w_x$  is symmetric and positive semidefinite and, using these two facts in (14), leads immediately to the standard Slutsky relations:<sup>8</sup> the symmetry of compensated price derivatives (or substitution terms)

$$(15) \quad \left. \frac{\partial x^j}{\partial p_k^l} \right|_{y_k-\text{comp}} = \left. \frac{\partial x^l}{\partial p_k^j} \right|_{y_k-\text{comp}}$$

and the nonpositivity of the compensated own-price derivatives,

$$(16) \quad \left. \frac{\partial x^j}{\partial p_k^j} \right|_{y_k-\text{comp}} \leq 0$$

We now wish to compare the substitution terms between two commodities (say commodity  $j$  and commodity  $l$ ) in two different income constraints. In other words, we would like to compare

$$\left. \frac{\partial x^j}{\partial p_k^l} \right|_{y_k-\text{comp}} \quad \text{with} \quad \left. \frac{\partial x^j}{\partial p_h^l} \right|_{y_h-\text{comp}}$$

For the former, we have equation (14), and for the latter we may write, using (14) and (15):

$$(17) \quad \left. \frac{\partial x^j}{\partial p_h^l} \right|_{y_h-\text{comp}} = \left. \frac{\partial x^l}{\partial p_h^j} \right|_{y_h-\text{comp}} \quad (\text{continued})$$

<sup>8</sup> The Slutsky matrix is of rank  $n-m$ , and this can be seen by considering the various derivatives directly. This fact has the implication that at least  $n-m$  of the compensated own-price derivatives are negative.

$$\left. \frac{\partial x^j}{\partial p_h^l} \right]_{y_h-\text{comp}} = - \frac{\partial^2 w_x / \partial p_h^l \partial p_h^j}{v_l}, \quad (21)$$

$$\frac{v_k^l}{v_k} = \frac{v_h^l}{v_h},$$

for  $t=1, \dots, m$ . By selecting  $i=h$  in (14) and  $t=k$  in (17), we arrive at the following result:

$$(18) \quad \frac{\left. \frac{\partial x^j}{\partial p_k^l} \right]_{y_k-\text{comp}}}{\left. \frac{\partial x^j}{\partial p_h^l} \right]_{y_h-\text{comp}}} = \frac{v_k}{v_h}$$

Since both  $v_k$  and  $v_h$  are positive, one may conclude from (18) that if two commodities are substitutes in one constraint, then they are substitutes in all constraints, and similarly in the case of complements.

In addition to income-compensated price changes, one may wish to investigate other types of compensated variations, such as income-compensated income changes or price-compensated price changes. Consider, for example, a change in the  $k$ th income, accompanied by a compensating change in the  $h$ th income. The rate of change of  $x^j$  resulting from this compensated variation is given by:

$$(19) \quad \left. \frac{\partial x^j}{\partial y_k} \right]_{y_h-\text{comp}} = \frac{\partial x^j}{\partial y_k} + \left. \frac{\partial y_h}{\partial y_k} \right]_{v \text{ const}} \frac{\partial x^j}{\partial y_h} \\ = \frac{\partial x^j}{\partial y_k} - \frac{v_k}{v_h} \frac{\partial x^j}{\partial y_h}$$

Now let us consider a change in the price  $p_k^l$ , accompanied by a compensating change in  $p_h^l$ . The effect of such a compensated change on  $x^j$  is given by:

$$(20) \quad \left. \frac{\partial x^j}{\partial p_k^l} \right]_{p_h^l-\text{comp}} = \frac{\partial x^j}{\partial p_k^l} + \left. \frac{\partial p_h^l}{\partial p_k^l} \right]_{v \text{ const}} \frac{\partial x^j}{\partial p_h^l} \\ = \frac{\partial x^j}{\partial p_k^l} - \frac{v_k^l}{v_h^l} \frac{\partial x^j}{\partial p_h^l}$$

But, from (10) we have that:

so that

$$(22) \quad \left. \frac{\partial x^j}{\partial p_k^l} \right]_{p_h^l-\text{comp}} = \frac{\partial x^j}{\partial p_k^l} - \frac{v_k}{v_h} \frac{\partial x^j}{\partial p_h^l}$$

From (19) and (22), with the aid of (18), one may derive a relationship between the effects of income-compensated income changes and the effects of price-compensated price changes. To this end, let us evaluate the following quantity:

$$\left. \frac{\partial x^j}{\partial p_k^l} \right]_{p_h^l-\text{comp}} + x^l \left. \frac{\partial x^j}{\partial y_k} \right]_{y_h-\text{comp}}$$

Using (19) and (22), we may rewrite this quantity as:

$$\frac{\partial x^j}{\partial p_k^l} + x^l \frac{\partial x^j}{\partial y_k} - \frac{v_k}{v_h} \left[ \frac{\partial x^j}{\partial p_h^l} + x^l \frac{\partial x^j}{\partial y_h} \right]$$

or as

$$\left. \frac{\partial x^j}{\partial p_k^l} \right]_{y_k-\text{comp}} - \frac{v_k}{v_h} \left[ \left. \frac{\partial x^j}{\partial p_h^l} \right]_{y_h-\text{comp}} \right]$$

and this last quantity is zero, by virtue of (18). So, we have the following result:

$$(23) \quad \left. \frac{\partial x^j}{\partial p_k^l} \right]_{p_h^l-\text{comp}} = - x^l \left. \frac{\partial x^j}{\partial y_k} \right]_{y_h-\text{comp}}$$

Thus, a change in prices that keeps utility constant is equivalent to a change in incomes (in the opposite direction) that keeps utility constant, weighted by quantity demanded. This suggests that the combination of price changes which keeps utility constant involves no substitution effects, only income effects. By increasing  $p_k^l$  we decrease utility and have a substitution effect away from good  $l$ . By decreasing  $p_h^l$  we increase utility and have a substitution effect toward good  $l$ . The relative size of changes in these two prices which keeps utility constant just balances the

substitution effects so that they net out, leaving only income effects.

### III. Back to the Uncertainty Setting

In Section I, we formulated a fairly general portfolio-selection problem for a consumer facing uncertainty. We found that this problem is isomorphic to the choice problem of a consumer operating in a world of rationing. The significance of this isomorphism lies in the fact that it permits the direct use of results from demand theory under rationing in a theory of investment and portfolio choice. The problem that we have been considering is:

$$\begin{array}{ll} \text{maximize } u(x) \\ \text{subject to: } & Px = y \end{array}$$

In the world of rationing,  $x$  is a vector whose components are quantities of various commodities demanded by the consumer. In the uncertainty setting,  $x$  is given by:

$$x = \begin{bmatrix} c_1 \\ \vdots \\ c_n \\ -s_1 \\ \vdots \\ -s_m \end{bmatrix}$$

The first  $n$  components of  $x$  are, once again, quantities demanded, while the last  $m$  components are the negative of the quantities invested in various risky securities. It is to these quantities that we now wish to apply the results outlined in Section II.

But before we can proceed to apply the results in Section II to the uncertainty model, the following technical issue must be settled. All the results in Section II depend on demand functions being single valued (differentiability, then, being assumed). This, in turn, depends on utility

being strictly quasiconcave. But the utility  $\hat{u}$  in equation (5), to which we would like to apply the results of rationing theory, is *not* strictly quasiconcave. Indeed, the investment levels  $s_1, \dots, s_m$  appear in it as perfect substitutes. This may lead to demand *correspondences* (i.e., many-valued demand functions) and thus vitiate the analogy that we are looking for. However, this difficulty is apparent rather than real, provided that the securities are linearly independent, i.e., the columns of the  $A$  matrix are linearly independent. Of course, this requires at least as many states of nature as securities. In fact, if the utility  $u$ , from which  $\hat{u}$  is derived, is strictly quasiconcave, then the security demands  $s_1, \dots, s_m$  will be single valued. This may be seen by using the constraints to eliminate  $c_1, \dots, c_n$  from  $\hat{u}$ . The result will be a function depending on  $s_1, \dots, s_m$  alone, and this function is easily shown to be strictly quasiconcave when the securities are linearly independent. Thus, optimal investment levels are single valued in the parameters.

Let us turn first to the Slutsky relations. Using the notation of Section I, with  $s_j$  standing for investment in the  $j$ th security and  $\alpha_{ij}$  standing for the gross rate of return on security  $j$  in state  $i$ , we may define the effect of an income-compensated change in yield (see (11)) as follows:

$$(24) \quad \left. \frac{\partial s_i}{\partial \alpha_{jk}} \right|_{y_j\text{-comp}} = \frac{\partial s_i}{\partial \alpha_{jk}} - s_k \frac{\partial s_i}{\partial y_j}$$

(Note that it is the *negative* of the investment level which takes the place of the quantity demanded in the rationing model. This fact accounts for the sign changes in the various formulas.) The standard Slutsky properties now take the following form:<sup>9</sup>

<sup>9</sup> Consideration of the matrix of compensated derivatives of  $s$ , with respect to  $\alpha$ , shows it to be of full rank. This implies the strict inequality in (26).

$$(25) \quad \left. \frac{\partial s_i}{\partial \alpha_{jk}} \right]_{y_j - \text{comp}} = \left. \frac{\partial s_k}{\partial \alpha_{ji}} \right]_{y_j - \text{comp}}$$

and

$$(26) \quad \left. \frac{\partial s_i}{\partial \alpha_{ji}} \right]_{y_j - \text{comp}} > 0$$

Equation (25) says that the effect on investment in security  $i$  of a compensated change in the yield of security  $k$  in state  $j$  equals the effect on investment in security  $k$  of a compensated change in the yield of security  $i$ , also in state  $j$ , where both compensations are made in terms of noninvestment income in state  $j$ . Equation (26) says that an increase (say) in the yield of security  $i$  in state  $j$ , compensated by a decrease in noninvestment income in state  $j$ , will result in an increase in the consumer's holdings of security  $i$ .

As in standard consumer theory, assumptions on the signs of income effects can be combined with the Slutsky conditions to obtain signs of noncompensated derivatives. As an example, let us consider the result that if the holding of a security is positive and increasing with initial wealth, and if there exists a riskless security, then the level of holding of the security increases with an increase in its yield in all states of nature.<sup>10</sup> (See Agnar Sandmo for this result.) Our assumption is that, for some  $i$ ,

$$(27) \quad s_i > 0, \quad \frac{\partial s_i}{\partial y_0} > 0$$

We wish to prove that

$$(28) \quad \sum_{j=1}^n \frac{\partial s_i}{\partial \alpha_{ji}} > 0$$

We shall prove a somewhat more general

<sup>10</sup> If we assign probabilities to states of nature, the yields of a security become a random variable. The change in yields which we are considering is then equivalent to a constant rightward shift of the distribution.

result, which implies (28) in the presence of a riskless asset.

Since optimal consumption depends only on consumption possibilities (i.e., on the budget set) and not on the particular combination of incomes and yields that gives rise to these consumption possibilities, we see that the response of  $c_i$  to a change in  $y_0$  is the same as its response to changes in  $y_1, \dots, y_n$  which satisfy  $dy_i = \alpha_{ik} dy_0$  for some  $k$ . This follows since the effect of these future income changes on consumption can be accomplished also by changing the holdings of the  $k$ th security after a change in present income. This implies that the responses of security holdings (except for the holdings of the  $k$ th security) to these two kinds of changes must also be the same in order to result in the same changes in consumption. In other words, we have, for  $i \neq k$ , that

$$(29) \quad \frac{\partial s_i}{\partial y_0} = \sum_{j=1}^n \alpha_{jk} \frac{\partial s_i}{\partial y_j}$$

Using (24), we may write

$$(30) \quad \sum_{j=1}^n \alpha_{jk} \left. \frac{\partial s_i}{\partial \alpha_{ji}} \right]_{y_j - \text{comp}} = \sum_{j=1}^n \alpha_{jk} \left. \frac{\partial s_i}{\partial \alpha_{ji}} \right]_{y_j - \text{comp}} + s_i \sum_{j=1}^n \alpha_{jk} \frac{\partial s_i}{\partial y_j}$$

By (26), the first term on the right-hand side of (30) is positive, provided that the yields  $\alpha_{1k}, \dots, \alpha_{nk}$  are nonnegative and not all zero. (This requirement is satisfied with limited liability, or if the  $k$ th security is riskless.) By (27) and (29), the second term on the right-hand side of (30) is also positive, provided  $i \neq k$ . If the  $k$ th security is riskless, so that  $\alpha_{jk}$  is the same for all  $j$ , then the desired result, (28), follows.

Since equation (29) does not hold for  $i = k$ , the analysis that we have made for the same absolute increase in all yields of a security cannot also be made for a pro-

portional increase in the yields of a given asset; for this requires evaluating

$$\sum_{j=1}^n \alpha_{ji} \frac{\partial s_i}{\partial \alpha_j}$$

To consider this case, let us introduce a shift parameter,  $\epsilon$ , and assume that the yields of the first security satisfy

$$\alpha_{i1}(\epsilon) = \epsilon \beta_i$$

for some constants  $\beta_1, \dots, \beta_n$ .<sup>11</sup> Let us define a new variable,  $z$ , by

$$z = \epsilon s_1$$

We can now restate the optimization problem as

$$\text{maximize } u(c_0, c_1, \dots, c_n)$$

subject to:

$$y_0 = c_0 + \epsilon^{-1}z + \sum_{j=2}^m s_j$$

$$c_i = y_i + \beta_i z + \sum_{j=2}^m \alpha_{ij} s_j, \quad i = 1, \dots, n$$

We can now consider changes in  $\epsilon$  compensated by changes in  $y_0$ . The Slutsky inequality for this type of change is given by

$$(31) \quad \left. \frac{\partial z}{\partial \epsilon} \right|_{y_0-\text{comp}} = \frac{\partial z}{\partial \epsilon} - \frac{z}{\epsilon^2} \frac{\partial z}{\partial y_0} > 0$$

Differentiating the definition of  $z$ , substituting into (31), and then setting  $\epsilon=1$ , we have

$$(32) \quad \left. \frac{\partial z}{\partial \epsilon} \right|_{y_0-\text{comp}} = \frac{\partial s_1}{\partial \epsilon} + s_1 \left[ 1 - \frac{\partial s_1}{\partial y_0} \right] > 0$$

Thus, with a positive income derivative of the first security, a compensated proportional increase in the yields of the first security leads to an increase in  $z$  but may lead to a decrease in  $s_1$ . Only when  $s_1$  is

sufficiently small, does the increase in returns necessarily lead to an increase in security holding.<sup>12</sup> This result may be explained more simply if we consider the case where the first security is riskless; that is,  $\alpha_{i1}(\epsilon) = \epsilon$ . Then,  $z$  is the quantity of second-period consumption which is purchased from the holdings of the first security. This physical quantity obeys the usual Slutsky rule. On the other hand,  $s_1$  is the *value* of the consumption purchased from the holdings of the first security. As in standard consumer theory, if the price of a commodity increases, the expenditure (in value terms) on this commodity may rise or fall even though the quantity purchased falls.

Turning now to equation (23) in Section II, we find, upon transforming it into the present framework, that

$$(33) \quad \left. \frac{\partial s_i}{\partial \alpha_{kj}} \right|_{\alpha_{hj}-\text{comp}} = s_j \left. \frac{\partial s_i}{\partial y_k} \right|_{y_k-\text{comp}}$$

Compensated yield changes are equivalent to weighted compensated income changes. Consider, for example, a utility-preserving change in the yields of security  $j$  in states  $k$  and  $h$ . Consider, also, a utility-preserving change in noninvestment income in states  $k$  and  $h$ . The effects of these two changes on investment in any given security are of the same sign if security  $j$  is purchased, and they are of opposite signs if security  $j$  is sold short. Also, compensated yield changes of a security that is neither purchased nor sold short do not affect investment behavior. From this it follows, for example, that there does not exist a constant-utility path that leads, via changes in the yields of a security, from purchases to short sales, or vice versa.

In terms of the substitution-effect argument made in the previous section, we can

<sup>11</sup> If probabilities are assigned to state of nature, then with limited liability, an increase in  $\epsilon$  is equivalent to a proportional rightward shift of the distribution function.

<sup>12</sup> A similar conclusion was reached by Tobin (1958) although his explanation of this result is different from ours. This result assumes that optimal consumption is in the interior of the consumption possibility set.

say that changes in the yields of a security that leave utility constant have no substitution effects.<sup>13</sup>

#### IV. A Single Security

The above results depend only on the convexity of preferences and the maximization of utility. In the case of a single security, we have obtained a further result that also depends solely on convexity. Let us begin by restating the maximization problem in this case, dropping the subscript one referring to this security:

$$(34) \quad \underset{s}{\text{Maximize}} \quad u(y_0 - s, y_1 + \alpha_1 s, y_2 + \alpha_2 s, \dots, y_n + \alpha_n s)$$

The first-order condition for this maximization is:

$$(35) \quad \sum_{i=1}^n \alpha_i u_i - u_0 = 0$$

where subscripts on  $u$  refer to partial derivatives. If we consider a single point,  $c^*$ , in consumption space, we see that this point will be the optimal consumption point chosen for a variety of initial conditions  $y$  and  $\alpha$ . In particular, given  $\alpha_3, \dots, \alpha_n, y_0, y_3, \dots, y_n$ , and  $s$ , for any level of  $\alpha_1$  we can find levels of  $\alpha_2, y_1$ , and  $y_2$  (although not necessarily nonnegative), so that  $c^*$  will still be chosen. From the levels of  $c^*$

<sup>13</sup> It is possible to state a risk-aversion axiom, analogous to the axiom proposed by Hayne Leland, leading to the result that, in the case of a single riskless security, the response of investment to income-compensated income changes is of determined sign. Equation (33) can then be used to obtain information on the response of investment to other parameter changes.

and the first-order condition, we see that these must satisfy

$$(36) \quad \alpha_2 u_2 = u_0 - \alpha_1 u_1 - \sum_{i=3}^n \alpha_i u_i,$$

$$y_1 = c_1^* - \alpha_1 s, \quad y_2 = c_2^* - \alpha_2 s$$

Returning to Figure 1, we see an example of two budget lines which give rise to the same optimal consumption point. Naturally, the response of  $s$  to changes in the parameters will depend, in general, on which set of parameters has resulted in the choice of  $c^*$  as well as the point  $c^*$  itself. However, from convexity we can obtain some information on the relationship among these derivatives at some of the parameter values.

The derivative we shall examine is  $\partial s / \partial \alpha_1$   $_{\alpha_2-\text{comp}}$ . To calculate this, let us first calculate  $\partial s / \partial \alpha_k$  by differentiating the first-order condition:

$$(37) \quad \frac{\partial s}{\partial \alpha_k} = \frac{u_k + s \sum_{i=1}^n \alpha_i u_{ik} - s u_{0k}}{\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j u_{ij} - \sum_{i=1}^n \alpha_i u_{0i} + u_{00}}$$

We can now write the compensated derivative from (22), using the special structure of this problem which implies  $v_i = u_i$ . We obtain equation (38). The denominator of this expression is negative, given convexity. Thus, the sign of the derivative is the same as the sign of the numerator. If

$$(38) \quad \left. \frac{\partial s}{\partial \alpha_1} \right)_{\alpha_2-\text{comp}} = - \frac{s \sum_{i=1}^n \alpha_i \left( u_{i1} - \frac{u_1}{u_2} u_{i2} \right) - s \left( u_{01} - \frac{u_1}{u_2} u_{02} \right)}{\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j u_{ij} - \sum_{i=1}^n \alpha_i u_{0i} + u_{00}}$$

we think of the derivative as a function of the parameters that lead to the same optimal  $c^*$ , the derivatives of  $u$  are independent of changes in  $\alpha_1$  and  $\alpha_2$ . If we consider changes in  $\alpha_1$  and  $\alpha_2$ , satisfying (36), we can substitute in the numerator of (38) for  $\alpha_2$  from (36). Doing this, we obtain an expression which is linear in  $\alpha_1$ :

$$\begin{aligned}
 (39) \quad & s \left\{ \alpha_1 \left[ u_{11} - \frac{u_1}{u_2} u_{12} \right] \right. \\
 & + \left[ \frac{u_0}{u_2} - \alpha_1 \frac{u_1}{u_2} - \sum_{i=3}^n \alpha_i \frac{u_i}{u_2} \right] \\
 & \cdot \left[ u_{21} - \frac{u_1}{u_2} u_{22} \right] \\
 & + \sum_{i=3}^n \alpha_i \left[ u_{i1} - \frac{u_1}{u_2} u_{i2} \right] \\
 & \left. - u_{01} + \frac{u_1}{u_2} u_{02} \right\}
 \end{aligned}$$

Thus, assuming  $s > 0$ , there will be a unique pair  $(\alpha_1^*, \alpha_2^*)$ , such that  $\partial s / \partial \alpha_1)_{\alpha_2 = \text{comp}}$  is zero. For higher levels of  $\alpha_1$  (and lower levels of  $\alpha_2$ ), the derivative is negative, and vice versa. This sign result can be seen by differentiating (39) with respect to  $\alpha_1$ . This derivative then equals

$$\begin{aligned}
 (40) \quad & s \left( u_{11} - \frac{u_1}{u_2} u_{12} \right) \\
 & - s \frac{u_1}{u_2} \left( u_{21} - \frac{u_1}{u_2} u_{22} \right)
 \end{aligned}$$

This expression is negative by convexity and the assumed positive level of savings.

Let us restate this result. Consider a set of individuals with the same preferences and the same level of savings and consumption arising from choosing securities differing in their yields in states one and two. For those with a security showing a sufficiently high return in state 1, savings

decrease with further increases in the return in state 1, compensated by decreases in the return in state 2.

Now let us consider the situation when  $s = 0$ . Then,  $\partial s / \partial \alpha_1)_{\alpha_2 = \text{comp}}$  is zero at any of the pairs  $(\alpha_1, \alpha_2)$  which lead to  $c^*$ . (Note that  $y_1$  and  $y_2$  do not change with  $\alpha_1$  in this case.) This was the situation pointed out near the end of the preceding section. For this special case, we can examine the geometry of this result by returning to Figure 1. If savings are zero, the intersection of the two lines represents the initial position as well as the optimal consumption point. Compensated changes in  $\alpha_1$  and  $\alpha_2$  are rotations of the budget line in the tangent plane and so preserve the property that the initial position is the optimal consumption point and so security holdings are zero.

#### REFERENCES

- K. Borch, "A Note on Uncertainty and Indifference Curves," *Rev. Econ. Stud.*, Jan. 1969, 36, 1-4.
- M. S. Feldstein, "Mean-Variance Analysis in the Theory of Liquidity Preference and Portfolio Selection," *Rev. Econ. Stud.*, Jan. 1969, 36, 5-12.
- H. Leland, "Saving and Uncertainty: The Precautionary Demand for Saving," *Quart. J. Econ.*, Aug. 1968, 82, 465-73.
- R. Roy, *De l'Utilité, Contribution à la Théorie de Choix*, Paris 1943.
- P. A. Samuelson, *Foundations of Economic Analysis*, Cambridge 1947.
- A. Sandmo, "Capital Risk, Consumption, and Portfolio Choice," *Econometrica*, Oct. 1969, 37, 586-99.
- J. Tobin, "A Survey of the Theory of Rationing," *Econometrica*, Oct. 1952, 20, 521-53.
- , "Liquidity Preference as Behavior Toward Risk," *Rev. Econ. Stud.*, Feb. 1958, 25, 65-86.
- , "Comment on Borch and Feldstein," *Rev. Econ. Stud.*, Jan. 1969, 36, 13-14.

# Life Cycle Saving: Theory and Fact

By KEIZO NAGATANI\*

"Our present behavior can only be affected by the expected future,—not the future as it will turn out but the future as it appears to us beforehand through the veil of the unknown."

*Irving Fisher*

The standard results of the life cycle theory of saving are: a) the level of consumption at any point of time depends on the present value of the entire lifetime earnings; and b) the proportionate rate of change of the marginal utility of consumption at any point of time is equal to the difference between the subjective discount rate and the objective discount rate or the rate of interest (see Menahem Yaari (1964)). This is for the case of perfect certainty. Yaari (1965) and James Tobin studied the case in which individuals are subject to uncertainty with respect to the duration of their lifetime but in which a competitive insurance-annuity scheme is available. The result for this case is, to no surprise, the same as that for the perfect certainty case, except that the discount rates, both subjective and objective, are now raised to incorporate such biological uncertainty. The competitive insurance premium or annuity yield is equal to the instantaneous probability of death times the value of the current net worth. To the extent that such payments or receipts take place, the present value of lifetime earnings will differ from that in the perfect certainty case, but the *time shape* of optimal consumption remains unaffected.

\* Associate professor of economics, University of British Columbia. I am grateful to John Cragg, Rodney Dobell, Wayne Thirsk, and Robert Jones for valuable comments and suggestions for an earlier draft of this paper. My thanks are also due to the editor and a referee of this *Review* for their helpful advice. All errors and shortcomings are of course mine.

The major implication of these models is that the lifetime allocation of consumption is independent of the timing of income realization. But in a more general case with economic uncertainty, the timing of income realization and the degree of uncertainty associated with future incomes will be the important determinants of the optimal consumption plan. As is well known, these matters were discussed at some length by Irving Fisher. Recent formulations of the consumer's lifetime allocation problem in the stochastic dynamic programming framework are attempts to formalize Fisher's "Third Approximation," (see, for example, Nils Hakkanson). While the model presented below follows this approach in spirit, a simplifying device is employed to facilitate computations.

The available household data, on the other hand, show that the time shape of consumption typically has a "hump" around the middle years of life (see Table 4 below). The cause for such a hump, however, has not so far been established. The family size may be the major cause (Tobin), but it has been found rather insignificant (see H. Watts). While the above standard model has been applied extensively to the estimation of an aggregate consumption function, its micro evidence is rather weak. Thus there is room for an improvement.

The first and main purpose of this paper is to present a simple model of life cycle saving which builds on the Yaari-Tobin result by incorporating uncertainty concerning future human incomes.<sup>1</sup> Such uncertainty may be due to the general lack of

<sup>1</sup> Allowing for uncertainty about future interest rates poses no difficulty. But we shall abstract from it for ease of exposition.

knowledge about future developments of the environment, but is mainly due to the individual's limited ability to assess his own potential, luck, etc. While the individual can only have imperfect knowledge about the future, it seems reasonable to expect that uncertainty about the future will be reduced to some extent by his own experience. To the extent that experience changes his future outlook, his subjective evaluation of future earning power will change, and hence, his optimal consumption plan will be revised. The model attempts to interpret the observed consumption profile as one such continually revised optimal consumption profile. It will be found that both the time shape of human income and the uncertainty associated with it are crucial in determining the time shape of the optimal consumption.

The second purpose is to reexamine the notion of optimal consumption profile as Lester Thurow employed in a recent issue of this *Review*. Thurow's optimal consumption profile is essentially that income profile at which savings are zero. We shall argue, on the basis of our model, that such a notion is questionable, because (a) the very notion of optimality depends on the time shape of human income, and (b) such an income profile would not free individuals from the burden of uncertainty.

### I. The Model

We consider a hypothetical individual at age  $v(0 \leq v < T)$  where  $[0, T]$  is his lifetime assumed to be known in advance.<sup>2</sup> He attempts to reallocate his earnings to maximize the lifetime utility of consumption. For simplicity, we assume both the initial and the terminal wealth to be zero;  $a(0) =$

$a(T) = 0$ . We also assume that he possesses a utility function  $u[c(t)]$  which is time-invariant and concave everywhere. His net worth  $a(v)$  is given as a datum by his past behavior. Besides, he expects an uncertain flow of human income  $w_v(t)$ , ( $v \leq t \leq T$ ). The subscript  $v$  is attached to emphasize that the uncertainty or the probability properties of the random variable  $w(t)$  for any future date  $t$  depends on the date of evaluation. This is because such probability notion is essentially subjective and is conditioned by the individual's past experience. His problem is to plan his consumption for the rest of his life in such a way that the sum of the utility of consumption is maximal subject to a relevant wealth constraint. It remains to investigate what the wealth constraint is in the presence of such uncertainty.

If there is no uncertainty, the wealth constraint faced by an individual of age  $v$  is simply

$$\begin{aligned} a(v) + \int_v^T w(\tau) \exp \left\{ - \int_v^\tau r(x) dx \right\} d\tau \\ = \int_v^T c(\tau) \exp \left\{ - \int_v^\tau r(x) dx \right\} d\tau \end{aligned}$$

That is, the present value of future consumption (right-hand side) is equal to the sum of the current net worth and the present value of future human income (left-hand side). When there is uncertainty about future human income, one is tempted to simply replace  $w(\tau)$  by its expected value  $\bar{w}_v(\tau)$ . While this may be a permissible behavior, it implies a strong, and rather unrealistic, assumption that the burden of risk associated with current income is entirely born by current saving or future consumption. More realistically, the burden of such risk will be shared by current consumption and current saving. To capture this idea, we choose to write the wealth constraint as

<sup>2</sup> When lifetime is uncertain,  $T$  becomes a stochastic variable. In such a case, interpret our  $T$  as the lowest value of  $T$  for which the probability of survival to that age is zero. Since a competitive insurance-annuity scheme eliminates the effect of such uncertainty, we shall ignore this source of uncertainty.

$$\begin{aligned}
 (1) \quad & a(v) + \int_v^T \bar{w}_v(\tau) \cdot [1 - z_v(\tau)] \\
 & \cdot \exp \left\{ - \int_v^\tau r(x) dx \right\} d\tau \\
 & = \int_v^T c_v(\tau) \exp \left\{ - \int_v^\tau r(x) dx \right\} d\tau
 \end{aligned}$$

where  $z_v(\tau)$  denotes the risk measure per dollar associated with  $w_v(\tau)$ , ( $v \leq \tau \leq T$ ).

In order for this procedure to be meaningful, we need to show that  $1 > z_v(\tau) > 0$  whenever  $w_v(\tau)$  is uncertain. It turns out that a necessary and sufficient condition for this to be true for any  $w_v(\tau)$  is that, besides the concavity, the utility function has a positive third derivative, i.e., the marginal utility curve be convex from below.<sup>3</sup> Fortunately, this is not a demand-

<sup>3</sup> A sketch of the proof is as follows. First  $1 - z_v(t) > 0$  is intuitively clear, when  $w_v(t)$  is positive. Second to show that  $z_v(t) > 0$ , consider a two-period maximization problem where the only source of income  $y$  is to be realized in the second period and is random. Assume in particular that the income in the second period will be  $y'$  with probability  $p$  and  $y''$  with probability  $(1-p)$ . The expected value is  $\bar{y} = py' + (1-p)y''$ . Varying degree of uncertainty can be represented by variations in  $y'$  and  $y''$  with fixed  $p$  and fixed  $\bar{y}$ . Assume  $y' < y''$ . The problem is now to maximize  $u(c_1) + pu(y' - c_1) + (1-p)u(y'' - c_1)$ , with respect to the choice of  $c_1$ , the consumption in the first period. The first-order condition is

$$u'(c_1) - pu'(y' - c_1) - (1-p)u'(y'' - c_1) = 0$$

To evaluate the effect on  $c_1$  of a change in the degree of uncertainty, we constrain the variations of  $y'$  and  $y''$  by  $0 = d\bar{y} = pdy' + (1-p)dy''$ , and calculate

$$dc_1/dy'' = \frac{(1-p)[u''(y'' - c_1) - u''(y' - c_1)]}{u''(c_1) + pu''(y' - c_1) + (1-p)u''(y'' - c_1)}$$

If an increase in uncertainty means a smaller certain income,  $dc_1/dy'' < 0$ . Under the assumption of strict concavity of  $u$ , this will be the case if, and only if,  $u''(y'' - c_1) - u''(y' - c_1) > 0$ , i.e., the  $u'$  curve is convex from below. For a general distribution of  $y$ , it is easiest to compare the certainty and the uncertainty cases. The first-order condition for the certainty case is

$$(a) \quad u'(\bar{c}_1) = u'(\bar{y} - \bar{c}_1) \quad \text{or} \quad u'(\bar{c}_1) = u'[E(y - \bar{c}_1)]$$

For the uncertainty case, the first-order condition is

$$(b) \quad u'(c_1) = E[u'(y - c_1)]$$

If  $u'(\cdot)$  is strictly convex,  $E[u'(y - c_1)] > u'[E(y - \bar{c}_1)]$  at  $c_1 = \bar{c}_1$ , i.e., the right-hand side of (b)  $>$  the right-hand side of (a). But at  $c_1 = \bar{c}_1$ , the left-hand side of (a)

ing condition. If the individual is non-satiated with consumption, this condition must hold at least in the large. We might add that all the popular utility functions employed in similar stochastic models satisfy this condition, with an obvious exception of the quadratic function. Given this condition, "the risky remote income acts as the equivalent of a smaller remote income," which Fisher thought was the normal case (p. 217).

We wish to make an additional assumption concerning the way the risk per dollar  $z_v(t)$  is related to the future date  $t$ . Specifically we assume that the risk is the greater, the more distant the future date, i.e.,  $z_v(t)$  is an increasing function of  $t$  for given  $v$ . As Fisher put it, "the risk applies more especially to the remoter income than to the immediate" (p. 217). These two assumptions can be compactly expressed by a risk premium of the form.

$$1 - z_v(t) = \exp \left\{ - \int_v^t \theta_v(\tau) d\tau \right\}$$

as applied to  $w_v(t)$ , where  $\theta_v(\tau) > 0$  for all  $v \leq \tau \leq T$ . Thus equation (1) is equivalent to

$$\begin{aligned}
 (2) \quad & a(v) + \int_v^T \bar{w}_v(\tau) \\
 & \cdot \exp \left\{ - \int_v^\tau [r(x) + \theta_v(x)] dx \right\} d\tau \\
 & = \int_v^T c_v(\tau) \exp \left\{ - \int_v^\tau r(x) dx \right\} d\tau
 \end{aligned}$$

= the left-hand side of (b). To restore equation (b), therefore,  $c_1$  must fall below  $\bar{c}_1$ , the optimal consumption for the certainty case. Since

$$\begin{aligned}
 E[u'(y - c_1)] - u'[E(y - c_1)] \\
 \simeq \frac{u''[E(y - c_1)]}{2} \{ (y - c_1) - E(y - c_1) \}^2
 \end{aligned}$$

when  $u'(\cdot)$  is convex, an increase in uncertainty in the sense of a greater variance of  $y$  ( $\bar{y}$  held constant) leads to a larger fall in  $c_1$ . It is interesting to note that the effect of the variance on consumption depends on the third, as well as the second, derivative of the utility function.

The problem can now be stated formally. The individual maximizes

$$(3) \quad \int_v^T u[c_v(\tau)] \exp \left\{ - \int_v^{\tau} \delta(x) dx \right\} d\tau$$

subject to the wealth constraint (2), given  $a(v)$ , and the nonnegativity constraint  $c_v(\tau) \geq 0$  for all  $v \leq \tau \leq T$ .

Assuming that the nonnegativity constraint is not binding, the optimal solution takes the form

$$(4) \quad u'[c_v(\tau)] = u'[c_v(v)] \cdot \exp \left\{ \int_v^{\tau} [\delta(x) - r(x)] dx \right\}$$

or

$$(4') \quad D \ln u'[c_v(\tau)] = \delta(\tau) - r(\tau); \quad (v \leq \tau \leq T),$$

where  $D \equiv d/d\tau$ . This is nothing but the standard result mentioned at the beginning. This consumption program would actually be followed for the rest of his life, if the individual felt no need to change it. The crucial point, however, is that  $c_v(v)$  depends on the individual's future outlook as of age  $v$ . Since his outlook as of any age is conditioned by his past experience,  $c_v(v)$  will be generally different from  $c_z(v)$ , ( $0 \leq z < v$ ), the consumption plan laid down earlier for the same age  $v$ . For example, the individual at age  $v$  may have found himself more successful in his career than he thought at age  $z$ . This will probably induce him to revise both his future earning power and future desired consumption upward. If such revisions do take place,  $c_z(v)$ ,  $z < v$ , will be abandoned in favor of  $c_v(v)$ , as the consumption plan effective at age  $v$ . Thus the optimal lifetime consumption plan generated by this model keeps shifting up and down with age among a family of parallel curves each with a different "intercept" representing the varying assessment of future earning power but with a common "slope" representing the term

$[\delta(\tau) - r(\tau)]$ . The figures are drawn to visualize the determination of an optimal lifetime consumption according to our model. In Figure 1, the marginal utility

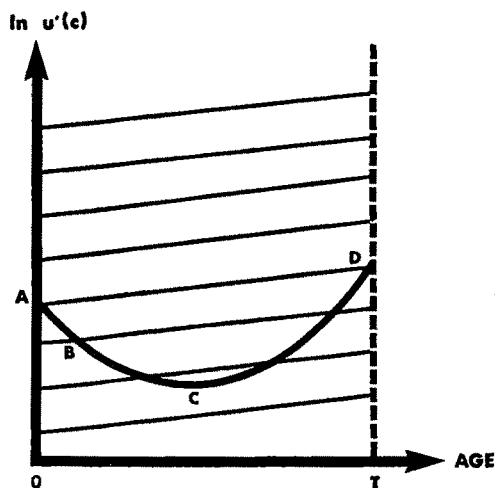


FIGURE 1

of consumption is measured along the vertical axis in logarithmic scale, and the age along the horizontal axis. A family of parallel curves (lines) represent fixed optimal plans with different levels of lifetime earnings. Each of these curves (lines) has a slope  $[\delta(v) - r(v)]$  at age  $v$ ,  $0 \leq v \leq T$ . Any one of these curves therefore can be interpreted as an optimal plan for the certainty case. In contrast, the present model generates an optimal plan as shown by a curve  $ABCD$ . It starts at age 0 at the point  $A$  on the basis of the individual's assessment of his future earning power at age 0. When he has lived for some years, part of the original future has turned into past in a way which is generally different from the original expectation. Even if his experience has been consistent with his original expectation, he will now have a firmer idea about future, and his current assessment of his future earning power will be different from what he originally thought. This change in outlook induces him to shift to the point  $B$ , say, on a different curve (line). In Figure 1,

such revisions are assumed to take place continuously to yield a smooth curve. Obviously, the shape of the curve like  $ABCD$  can only be determined when all the information about the pattern of change of the individual's outlook over time has been known, in addition to the complete knowledge of the utility function and the discount rates. In Figure 1, the curve  $ABCD$  is drawn to approximate a realistic consumption profile. The translation of this curve into a consumption curve is shown in Figure 2. In fact, our purpose is to interpret actual household behavior as one such optimal solution.

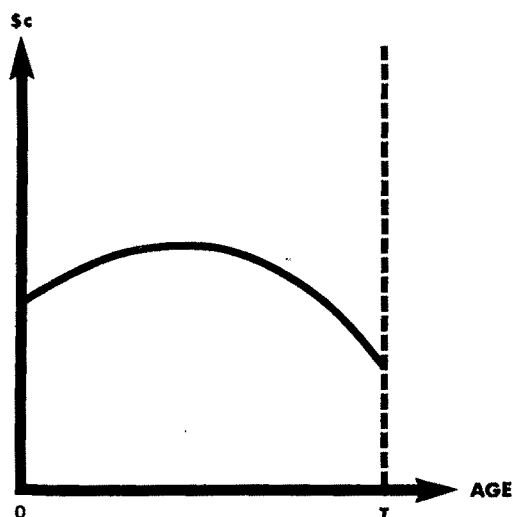


FIGURE 2

To summarize: Under some reasonable assumptions on the utility function, uncertainty about future income is translated into a risk premium which adds to the market interest rate in discounting future income. At any age, the individual spells out the optimal consumption plan according to the standard theory of life cycle saving. But since realized income is generally different from what he expected and since his future outlook (the expected value as well as the degree of uncertainty associated with future income) is condi-

tioned by his age, the wealth constraint itself becomes a function of his age. As his future outlook changes, therefore, the wealth constraint changes urging him to revise his consumption plan. While the revision does not involve a change in his basic behavioral rule, it involves a shift from one optimal profile to another in response to the change in the wealth constraint. It remains to study the pattern according to which such shifts take place.

## II. The Behavior of an Average Household

The purpose of this section is to apply the model presented in Section I to an "average" household, or the average behavior of a large number of households. The reasons for confining ourselves to an average household are (a) that the average behavior is sufficient to draw macroeconomic implications of the model, and (b) it justifies some simplifying assumptions concerning the manner in which future outlook changes with age.

In regard to the behavior of such an average household, we make the following three assumptions. First, all the expected future incomes  $\bar{w}_v(t)$ ,  $v \leq t \leq T$ , are independent of  $v$ , the age at which they are evaluated, and furthermore, that they are the same as their realized values,  $w(t)$ . In other words, the average household has the correct expectation of the mean values of future incomes. Needless to say, this does *not* mean that it has perfect foresight. Uncertainty is still with it. Second, for a similar reason, the risk premium  $\theta_v(t)$  is independent of the date of assessment. For individual households  $\theta_v(t)$  for any future date  $t$  may change in various ways with  $v$ . But for the average household, these changes may be taken to cancel out. Thus  $\theta$  is assumed to be some positive prescribed function of time. This assumption implies that the risk premium per dollar,  $z_v(t)$ , is a decreasing function of  $v$  for fixed  $t$ . Third, we assume that revisions of the

$$(5) \quad c_v(v) = \hat{c}(v) = \frac{a(v) + \int_v^T w(\tau) \exp \left\{ - \int_v^\tau [r(x) + \theta(x)] dx \right\} d\tau}{\int_v^T \exp \left\{ \frac{1}{k} \int_v^\tau [(1-k)r(x) - \delta(x)] dx \right\} d\tau}$$

consumption plan take place continuously. This assumption is mainly for technical reasons. In addition to these assumptions, we shall be assuming hereafter that the utility function is of the constant elasticity of marginal utility variety. This class of utility function can be written as

$$u(c) = \frac{c^{1-k}}{1-k}; \quad (k > 0); \quad \lim_{k \rightarrow 1} \frac{c^{1-k}}{1-k} = \ln c$$

Such utility functions satisfy all the conditions we need.

Given all these assumptions, we derive equation (5) from equations (2) and (4). By the assumption of continuous revisions, the plan laid down at age  $v$  is effective only at that age. Hence, we shall write  $c_v(v)$  as  $\hat{c}(v)$ . Since  $r + \theta > r$ , and since the *ex post* rate of return on wealth is only  $r$ , the past and the future incomes are no longer directly additive. The plan  $\hat{c}(v)$  therefore depends on the current net worth  $a(v)$ . The implications of the solution (5) become clearer if we differentiate it logarithmically with respect to  $v$ , noting the fact that the *ex post* rate of return on wealth is only  $r(v)$ . The result is

$$(6) \quad D \ln \hat{c}(v) = \frac{[r(v) - \delta(v)]/k}{+ [1 - \alpha(v)]\theta(v)}$$

where  $D \equiv d/dv$  and  $\alpha(v)$  is given by

$$(7) \quad \alpha(v) = a(v) / \left[ a(v) + \int_v^T w(\tau) \exp \left\{ - \int_v^\tau [r(x) + \theta(x)] dx \right\} d\tau \right]$$

which is the ratio of the realized nonhuman wealth to the total (human and non-human) wealth at any age  $v$ . Clearly  $\alpha(v)$  is equal to or less than unity, and may take on negative values. Since the term in the dotted rectangle of (6) is the certainty counterpart, the burden of uncertainty is shown by the term  $\theta(v) [1 - \alpha(v)]$ . When  $\theta(v) > 0$ , the time shape of optimal consumption becomes dependent on the risk premium  $\theta(v)$  and the ratio  $\alpha(v)$ . For given profiles of  $r$ ,  $\delta$ , and  $\theta$ ,  $\alpha(v)$  is primarily determined by the time shape of human income: if it concentrates in later years, the smaller the algebraic value of  $\alpha(v)$ , and hence, the greater the burden of uncertainty.

It is now quite easy to explain how the lifetime consumption profile of the average household is determined. Starting with the given initial net worth (say, zero), the consumption in the uncertainty case will begin and remain lower than the consumption in the certainty case for a while, because of the higher discount of future incomes, as shown in (5). But this means that if a common human income keeps flowing in, the net worth position will become higher than that in the certainty case. Under our assumptions that all the expected human incomes are realized and that the individual plans to leave no bequest, if the consumption for one case started low relative to the consumption for the other, there will be some intermediate time point at which the two consumptions equal each other and after which the relative consumption level will be reversed. In fact, equation (6) shows that the proportionate rate of change of consumption for

the uncertainty case is higher than that for the certainty case. Since the two consumptions will equal each other when the positive difference in the nonhuman wealth exactly offsets the negative difference in human wealth due to risk, such a break-even point will be reached the sooner (the later), the more concentrated the human income profile is in the early (late) years. Thus the optimal consumption profile in the present revision model becomes dependent on that of human income in such a manner that the former tends to resemble the latter.

The three tables are intended to illustrate such dependence of the consumption profile on the human income profile. The only fundamental difference among the three tables is the time shape of human income. In Table 1, the human income begins in the first year at \$1 and keeps rising to \$7 in the seventh year. The eighth through tenth years are the retirement. For this pattern of human income, con-

TABLE 1—DEPENDENCE OF CONSUMPTION ON HUMAN INCOME<sup>a</sup>

Year	$w(v)^b$	$\hat{c}(v)^c$	$w(v) - \hat{c}(v)$	$a(v)$
1	\$1	\$2.870	\$-1.870	\$ 0
2	2	2.877	-0.877	-1.945
3	3	2.914	0.086	-2.945
4	4	2.946	1.054	-2.974
5	5	2.973	2.027	-1.997
6	6	2.974	3.026	0.031
7	7	2.921	4.079	3.180
8	0	2.760	-2.760	7.549
9	0	2.609	-2.609	4.981
10	0	2.467	-2.467	2.467
11				0

<sup>a</sup>  $r = .04$ ,  $\theta = .06$ ,  $\delta = .10$  for all ages, and  $T = 10$  years.

<sup>b</sup> Both receipt of  $w$  and consumption are assumed to take place at beginning of each period. Net worth is measured also at beginning and before these transactions.

<sup>c</sup> Consumption figures are obtained from the discrete equivalent of equation (5), with  $k = 1$ .

$$\hat{c}(v) = \frac{\sum_{i=1}^{v-1} [w(i) - \hat{c}(i)](1+r)^{v-i} + \sum_{i=v}^{10} w(i)(1+r+\theta)^{v-i}}{\sum_{i=v}^{10} (1+\delta)^{v-i}}$$

TABLE 2—DEPENDENCE OF CONSUMPTION ON HUMAN INCOME<sup>a</sup>

Year	$w(v)$	$\hat{c}(v)$	$w(v) - \hat{c}(v)$	$a(v)$
1	\$7	\$3.454	\$ 3.546	\$ 0
2	6	3.438	2.562	3.688
3	5	3.374	1.626	6.506
4	4	3.284	0.716	8.457
5	3	3.170	-0.170	9.540
6	2	3.036	-1.036	9.745
7	1	2.886	-1.886	9.057
8	0	2.727	-2.727	7.458
9	0	2.578	-2.578	4.921
10	0	2.477	-2.477	2.477
11				0

<sup>a</sup> For underlying assumptions, see footnotes to Table 1.

sumption starts at \$2.870 in the first year and keeps rising till the sixth year when the maximum consumption is achieved at \$2.974. After that consumption gradually declines to \$2.467 in the tenth year. In Tables 2 and 3 more extreme human income profiles are assumed. In Table 2 the human income profile is reversed so that most of lifetime human income is realized in much earlier years. Reflecting the heavy concentration of human income in the early years, the consumption profile is also weighted toward the early years. Consumption shows a constant downward trend from \$3.454 to \$2.477. In Table 3, the human income profile is the same as in

TABLE 3—DEPENDENCE OF CONSUMPTION ON HUMAN INCOME<sup>a</sup>

Year	$w(v)$	$\hat{c}(v)$	$w(v) - \hat{c}(v)$	$a(v)$
1	\$0	\$2.157	\$-2.157	\$ 0
2	0	2.130	-2.130	-2.243
3	0	2.230	-2.230	-4.548
4	1	2.310	-1.310	-7.049
5	2	2.409	-0.409	-8.693
6	3	2.539	0.461	-9.467
7	4	2.693	1.307	-9.364
8	5	2.873	2.127	-8.380
9	6	3.070	2.930	-6.503
10	7	3.284	3.716	-3.716
11				0

<sup>a</sup> For underlying assumptions, see footnotes to Table 1.

Table 1, except that the retirement is replaced by an infancy. With increased burden of uncertainty, the consumption in Table 3 remains lower than the consumption in Table 1 in the first seven years but exceeding it in the last three years. In fact, the consumption in Table 3 almost always shows a rising trend due to negative net worth positions, in spite of the assumption that  $r - \delta < 0$ .<sup>4</sup>

To summarize: As was discussed in the previous section, consumption is locally influenced by the consumer's current net worth position in the presence of uncertainty. Under the few simplifying assumptions made for an average consumer, the revised consumption profile tends to resemble the profile of his human income.

### III. An Empirical Note

While the idea of revisions in consumption plans may sound attractive, it is not quite so easy to test this hypothesis directly against empirical data. Ideally we

need a survey which contains the lifetime record of the same households. In the absence of such data, however, we must be content with the available cross section data. Besides, equation (6) contains simply too many unknowns to estimate. The difficulty will remain even if we make a heroic assumption that the parameters  $k$ ,  $\delta$ , and  $\theta$  are all constant. The best we could do is to find a combination of these parameters that yields a closest fit to an observed average consumption profile.

In a recent issue of this *Review*, Thurow reports on his findings from the Bureau of Labor Statistics 1960-61 household budget data. The data contain information on the details of receipts and expenditures of seven age groups, each for different income brackets. The average expenditures for current consumption of different age groups are shown in column 3 of Table 4 below. The BLS data give the corresponding average figures for "money income after taxes" which contain both human and nonhuman incomes and are shown in column 1. The breakdowns of these total incomes between human and nonhuman incomes are not reported. We therefore

<sup>4</sup> Equation (6) implies that the *observed* rate of time preference is given by  $\bar{\delta}(v) = \delta(v) - k\theta(v)[1 - \alpha(v)]$  and hence  $\bar{\delta}(v)$  is generally less than  $\delta(v)$ :  $\bar{\delta}(v)$  is negatively related to  $\theta(v)$  and positively related to  $\alpha(v)$ .

TABLE 4—COMPARISONS OF ACTUAL, CALCULATED, AND OPTIMAL CONSUMPTION

Age	Money Income After Taxes (1)	Human <sup>a</sup> Income (2)	Actual Consump- tion (3)	Calculated Consump- tion <sup>b</sup> (4)	Optimal Consump- tion <sup>c</sup> (5)
Under 25	\$4,293	\$4,305	\$4,379	\$4,414	\$5,647
25-34	5,698	5,681	5,644	5,458	5,713
35-44	7,057	6,837	6,565	6,391	5,373
45-54	7,733	7,147	6,731	6,629	5,163
55-64	5,589	4,709	4,912	5,459	3,675
65-74	4,231	3,134	3,668	4,132	3,841
Over 75	3,016	1,805	2,712	2,946	2,162

<sup>a</sup> Human income (col. 2) was calculated by: (i) subtracting from col. 1  $(1.04)^{10}$  times the net worth under the assumption that the initial wealth is zero and (ii) by scaling down to equate its present value to that of actual consumption.

<sup>b</sup> Calculated consumption (col. 4) was obtained from our model by setting  $k=1$ ,  $r=.04$ ,  $\delta=.13$ ,  $\theta=.11$ , and using col. 2.

<sup>c</sup> Optimal consumption (col. 5) is Thurow's optimal consumption #1, excepting for a similar down-scaling.

calculated a human income profile in column 2. Column 3 is directly out of the data. According to column 3, actual consumption has a peak in the 45-54 years of age, tapering off toward both ends. In terms of our model, this is the direct consequence of a similarly humped human income profile. Column 4 reports on our experimental calculation based on our model and rather arbitrarily chosen values of the parameters. While already high values of  $\delta$  and  $\theta$  are assumed, the comparison with actual consumption suggests that  $\theta$  should be even higher (to make the hump larger) and that  $\delta$  should also be higher (to make the consumption profile more concentrated in the early years). Although we cannot have much faith in the accuracy of these estimates, it seems quite possible for our model to explain the observed consumption profile on the basis of the actual human income profile.

Thurrow's main purpose was to derive an "optimal" consumption profile from the same data. He did this by finding, for each age group, the level of money income after taxes at which saving is zero. As is seen from column 5, such an optimal profile is strongly weighted toward the young years. Comparing the actual and such optimal consumption profiles, Thurrow concludes that the actual lifetime pattern of income is a severe constraint on the desired lifetime distribution of consumption expenditures and that lifetime welfare levels might be substantially increased if the constraints on lifetime income redistribution could be lifted.

While this way of contrasting the actual result with an optimal one and attributing the gap to market imperfections is conventional, the validity of such a procedure is somewhat questionable. First, in terms of our model, the very notion of optimality is a relative concept. The actual consumption profile itself is an optimal profile relative to the given actual human income

profile. Any attempt to change the human income profile will therefore change the optimal consumption profile. Second, Thurrow's optimal consumption profile is characterized by zero saving. If we ignore initial assets, his optimal consumption corresponds to our revised consumption in which  $\alpha(v)=0$  for all  $v(0 \leq v \leq T)$ . From equation (6), it will then satisfy

$$(8) \quad D \ln c(v) = [r(v) - \delta(v)]/k - \theta(v)$$

which is still subject to uncertainty. In fact, the grand optimum would be attained when  $\theta \equiv 0$ . But this would mean that the individual would have to have his entire lifetime earnings in hand at the beginning of his life. The grand optimum consumption would be much more weighted toward the young years than Thurrow's optimum.

#### IV. Some Concluding Remarks

In this paper, I have attempted to provide a simple model of life cycle saving in an explicit context of uncertainty. The central result is that the lifetime consumption profile, instead of following a simple rule of growing at the rate  $[r - \delta]/k$ , will be generally dependent on the human income profile. The model seems capable of explaining and predicting the observed consumption profile as an optimal profile relative to the observed human income profile.

The key notion in the model is the continued revisions in the light of changing future outlook. In spirit, our model may be viewed as an example of the broader problem discussed by Robert Strotz.

The aggregate consumption function can be derived by integrating (5) over all the existing vintages of individuals. For a given pattern of lifetime human income, the discounted value of future human income can be written as a function of current human income. Therefore, the aggregate consumption will be a function of aggregate nonhuman wealth and wage

income, but it will not be a simple linear function of these variables. If we are to write the aggregate consumption as

$$C(t) = b_1A(t) + b_2W(t)$$

both  $b_1$  and  $b_2$  will depend on the age distribution of holders of wealth and recipients of wages and the demographic parameters, as well as on  $r$  and  $\delta$ . But at least our model provides a rationale for separating between the wealth and wage variables. The model also suggests that the income distribution among different age groups may be an important determinant of national saving behavior.

#### REFERENCES

- I. Fisher, *The Theory of Interest*, New York 1930.
- N. Hakkanson, "Optimal Investment and Consumption Strategies Under Risk for a Class of Utility Functions," *Econometrica*, Sept. 1970, 38, 587-607.
- R. H. Strotz, "Myopia and Inconsistency in Dynamic Utility Maximization," *Rev. Econ. Stud.*, June 1956, 23, 165-80.
- L. C. Thurow, "The Optimum Lifetime Distribution of Consumption Expenditures," *Amer. Econ. Rev.*, June 1969, 59, 324-30.
- J. Tobin, "Life Cycle Saving and Balanced Growth," in W. Fellner et al., eds, *Ten Economic Studies in the Tradition of Irving Fisher*, New York 1967.
- H. W. Watts, "Long-Run Income Expectations and Consumer Saving," in T. F. Dernberg, R. N. Rosett, H. W. Watts, eds., *Studies in Household Economic Behavior*, New Haven 1958.
- M. E. Yaari, "On the Consumer's Lifetime Allocation Process," *Int. Econ. Rev.*, Oct. 1964, 5, 304-17.
- , "Uncertain Lifetime, Life Insurance, and the Theory of the Consumer," *Rev. Econ. Stud.*, Apr. 1965, 32, 137-50.

# The Rationale of the Mean-Standard Deviation Analysis, Skewness Preference, and the Demand for Money

By S. C. TSIANG\*

In the January 1969 issue of the *Review of Economic Studies*, Karl Borch and Martin Feldstein separately criticize the widely used mean-variance analysis of portfolio selection. Borch contends that any system of upward sloping mean-standard deviation (henceforth abbreviated as  $E$ - $S$ ) indifference curves can be shown to be inconsistent with the basic axiom of choice under uncertainty. He points out that one can always pick any two points on an upward sloping indifference curve to represent two Bernoulli distributions with probability  $(1-p)$  of gaining the same amount  $\$x$  in both cases and probability  $p$  of gaining  $\$y_1$  in the case of one distribution and  $\$y_2$  in the case of the second distribution; with  $y_2 > y_1$ , if the second distribution represents the point to the northeast of the other point. Then by the dominance axiom, the second distribution must be preferred to the first, which contradicts the basic meaning of indifference curves. (See Figure 1.)

Feldstein, by using a *log* utility function and a lognormal distribution for investment outcome has shown that  $E$ - $S$  indifference curves for a risk-avertter need not be convex downwards, though upward sloping. They would change from convex to concave, once the standard deviation of the outcome exceeds the mean multiplied by  $1/\sqrt{2}$ . On the face of it, this seems

to suggest that risk aversion might decrease as risk itself is increased beyond a certain extent. Furthermore, he points out, as Paul Samuelson (1967) did before him, that in general it is not possible to define a preference ordering of portfolios of mixed investments in terms of  $E$  and  $S$  alone, since a linear combination of several stochastic variables would in general not be a variable of a two-parameter distribution, even if each of the constituent variables is of a two-parameter distribution (except the case where they are all normally distributed.)

This combined assault forced James Tobin, one of the pioneers of the  $E$ - $S$  analysis of portfolio choice, to acknowledge that it is applicable only if either the investor's utility function is quadratic, or if he regards the uncertain outcomes as all normally distributed (1969).

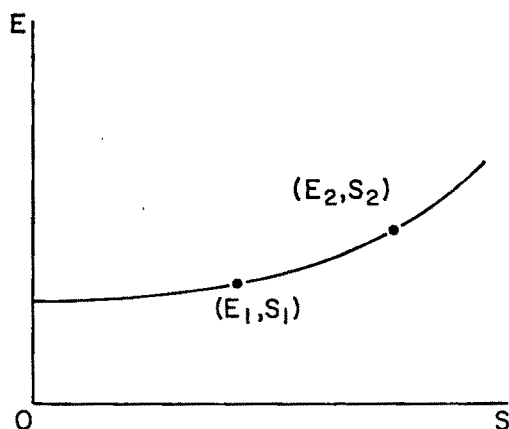


FIGURE 1

\* Professor of economics, Cornell University. I have benefited from comments by Paul Samuelson, Henry Wan and Leonard Mirman, who have read the draft, and from suggestions of the referee.

It is now generally recognized, however, that a quadratic function is not only limited in its range of applicability as a utility function, but that even within its range of applicability it involves the highly implausible implication of increasing absolute risk-aversion, which both Kenneth Arrow and John Hicks denounced as absurd. On the other hand, the assumption of normal distribution for all outcomes of risky investments and ventures is patently not realistic; for it would rule out all asymmetry or skewness in the probability distributions of returns. The study by Paul Cootner, for instance, suggests that returns of financial investments at least may be more likely to be distributed lognormally, rather than normally, being the cumulative products of random factors rather than their cumulative sums; and thus are likely to be positively skewed. Furthermore, progressive taxation, limited liability company organization, and hedging can also change otherwise symmetric distribution of returns of investments into skewed ones in net returns.

If the *E-S* analysis is restricted to either the case of quadratic utility function or the case of normally distributed investment outcomes, then it would be of very limited application indeed. The purpose of this paper is to point out that there is a justification for the use of the *E-S* analysis beyond the two cases to which Tobin would now confine its application, provided the aggregate risk taken by the individual concerned is small compared with his total wealth, including his physical, financial, as well as human wealth. It is not necessary that risk taken should be infinitesimally small in its absolute magnitude.<sup>1</sup>

However, it will also be pointed out that, because of the constraint on the slopes of *E-S* indifference curves indicated by our

discussion, the *E-S* analysis would not be capable of rationalizing the demand for idle cash in an investment portfolio, as it was originally called upon to do by Tobin (1958).

### I. Justification for Using Moments of Distributions for Preference Ordering of Uncertain Outcomes—Aversion to Dispersion and Preference for Skewness

Borch starts by pointing out that strictly speaking a consistent preference ordering of a set of uncertain outcomes of any different distributions can be established in terms of their respective first  $n$  moments only if the utility function of the individual concerned is a polynomial of degree  $n$ , as has been shown by Marcel Richter. Unfortunately, it is now generally recognized that polynomials are not suitable as utility function of wealth; for an appropriate utility function  $U(y)$  for a risk-averse individual, according to Arrow, should have the following essential properties.

(a)  $U'(y) > 0$ , i.e., marginal utility of wealth is positive;

(b)  $U''(y) < 0$ , i.e., marginal utility of wealth decreases with an increase of wealth;

(c)  $d[-U''(y)/U'(y)]/dy \leq 0$ , i.e., marginal absolute risk-aversion should, if anything, decrease with an increase in wealth;

(d)  $d[-yU''(y)/U'(y)]/dy \geq 0$ , i.e., marginal relative (proportional) risk-aversion should, if anything, increase with an increase in wealth.<sup>2</sup>

Polynomials as utility functions cannot satisfy these requirements at the same

<sup>1</sup> In this, our justification of the *E-S* analysis differs from Samuelson's (1970); for Samuelson's defense for the *E-S* analysis seems to rely upon the risk itself being very small in absolute magnitude.

<sup>2</sup> I do not see, however, any compelling reason why utility functions must be bounded both from above and from below. If Arrow's only concern is that "if the utility function is unbounded one can always construct an action with an infinite utility" (p. 25), I fail to see why we should be bothered with such possibilities. Samuelson has indicated to me in a correspondence that he agrees with me on the nonnecessity of the boundedness of utility functions.

time. Utility functions that satisfy these conditions, such as the negative exponential function  $U(y) = B(1 - e^{-ay})$ , and the family of constant elasticity utility functions:  $[1/(1-a)]y^{1-a}$ , ( $a > 0$ ), and  $U(y) = \log y$ ; etc., are not polynomials. However, nonpolynomials can generally be expanded into Taylor's series provided that they are continuous and have derivatives. That is, if  $y$  is a random variable, it can always be written as the sum of its mean plus the deviation from the mean, and, hence,

$$(1) \quad \begin{aligned} U(y) &= U(\bar{y} + h) = U(\bar{y}) + U'(\bar{y})h \\ &+ U''(\bar{y}) \frac{h^2}{2!} + U'''(\bar{y}) \frac{h^3}{3!} \\ &+ \dots + U^{(n-1)}(\bar{y}) \frac{h^{n-1}}{(n-1)!} \\ &+ R_n, \end{aligned}$$

where  $R_n = U^{(n)}(\bar{y} + \epsilon h) h^n / (n!)$ , ( $0 < \epsilon < 1$ ). The utility function thus becomes a polynomial in  $h$  (the deviation of  $y$  from its mean) with nonstochastic coefficients except for the remainder term. The expected utility is then

$$(2) \quad \begin{aligned} E[U(y)] &= \int_{-\infty}^{\infty} U(\bar{y} + h) f(h) dh \\ &= U(\bar{y}) + U''(\bar{y}) \frac{\bar{m}_2}{2} + U'''(\bar{y}) \frac{\bar{m}_3}{3!} \\ &+ \dots + U^{(n-1)}(\bar{y}) \frac{\bar{m}_{n-1}}{(n-1)!} \\ &+ \frac{1}{n!} E[U^{(n)}(\bar{y} + \epsilon h) h^n], \quad (0 < \epsilon < 1), \end{aligned}$$

where  $f(h)$  is the density function of  $h$ , the deviation of  $y$  from  $\bar{y}$ , and  $\bar{m}_2, \bar{m}_3, \dots, \bar{m}_{n-1}$  are the second, the third, and the successive higher central moments of the distribution. If this series can be shown to be convergent so that the remainder term can be neglected, then the expected utility can be treated as a func-

tion of the first  $(n-1)$  central moments of the distribution of  $y$  with constant coefficients as if the utility function were a  $(n-1)$ th order polynomial of  $y$ . The number  $n$  is to be varied to ensure sufficient degree of accuracy in approximation.

If the convergence of the series is sufficiently fast, so that, for fairly close approximation, the terms beyond the second moments can be neglected, then indeed the expected utility can be approximately determined by the first two moments, mean and variance, even if the utility function is not quadratic, and the uncertain outcomes not normally distributed. The crucial questions then are under what conditions we can expect the expansion of the utility function to converge quickly, and whether these conditions hold in the usual problems to which the *E-S* analysis is frequently applied.

Analysis of the expected utility function by its expansion is essentially the approach adopted by both Arrow and John Pratt in establishing the formulae for absolute and relative marginal risk aversion. In their respective works, they both dealt with the case of an individual, who, initially having no risk at all, is confronted with an infinitesimal and actuarially neutral risk. Since risk (variance) is assumed to be infinitesimally small, higher order central moments are assumed to be of even smaller orders and thus all omitted (see Pratt, p. 125). Then obviously expected utility may be approximated by

$$(2') \quad E[U(y)] = U(\bar{y}) + U'' \frac{S^2}{2},$$

from which it can be readily derived that

$$(2'') \quad \left. \frac{d\bar{y}}{dS} \right| = - \frac{SU''}{U'}$$

$$(E[U] = \text{constant})$$

One cannot, however, trace out a whole

indifference map for the mean ( $E=y$ ) and the standard deviation  $S$  on the assumption that the latter remains very, very small in absolute magnitude all the time. Fortunately, if we plug in some acceptable utility function into equation (2), it becomes clear that it is not really necessary for risk to remain very small in absolute magnitude. What is necessary for the  $E$ - $S$  analysis to be a good approximation is merely that risk should remain small *relatively* to the total wealth of the individual concerned.

One of the widely adopted functions for the utility of wealth is the negative exponential function, viz.,  $U(y) = B(1 - e^{-\alpha y})$ . It possesses all the four required properties of an acceptable utility function, except that the required property (c) is satisfied only marginally. With this function, absolute risk-aversion is invariant with an increase or decrease of wealth,<sup>3</sup> whereas it would probably conform more to empirical observations if absolute risk-aversion is decreasing with wealth.

Another widely used type of utility function of wealth is the family of constant elasticity utility functions of the form of  $U(y) = K + [1/(1-a)]y^{1-a}$ , ( $a > 0$ ), or the *log* function  $U(y) = \ln y$ . With this type of function, absolute risk-aversion would indeed be decreasing with wealth, but the relative risk-aversion is invariant with wealth. Compared with the former, however, this type of function has a serious defect in that it is generally either undefined or not real (imaginary), or yields negative marginal utility for zero or negative wealth.<sup>4</sup> Thus it should not be

<sup>3</sup> In this case, the absolute risk aversion  $= -U''/U' = \alpha$ .

<sup>4</sup> The relative risk aversion in the case of constant elasticity utility function of the form  $K + [1/(1-a)]y^{1-a}$  is  $yU''' / U' = a$ . In the case of *log* utility function, it is unity.

When  $a > 1$ ,  $U(y) = K + [1/(1-a)]y^{1-a}$ , like  $U(y) = \log y$ , approaches  $-\infty$ , as  $y \rightarrow 0$ . The utility function cannot be continuous when extended to the range of negative wealth.

used to deal with decisions in face of risk that might involve bankruptcy or negative wealth.

Since the most commonly observed pattern of behavior towards risk of a risk-averse individual is probably decreasing absolute risk-aversion coupled with increasing relative risk-aversion when his wealth increases, the ideal utility function of wealth should lie probably somewhere in between the negative exponential function and the constant elasticity function, and what is found to be true for both the exponential and the constant elasticity utility functions should be valid for most acceptable utility functions.

Let us first plug in the negative exponential utility function into equations (1) and (2) above. Writing  $y$  as  $(\bar{y} + h)$ , equation (1) becomes

$$(3) \quad U(y) = B - Be^{-\alpha y} \\ \cdot \left[ 1 - \alpha h + \frac{\alpha^2 h^2}{2!} - \frac{\alpha^3 h^3}{3!} + \dots \right]$$

which converges for all values of  $h$ , and the remainder term can always be neglected provided we extend the series to sufficient number of terms. Equation (2) then becomes

$$(4) \quad E[U(y)] = B - Be^{-\alpha \bar{y}} \\ \cdot \left[ 1 + \frac{\alpha^2 S^2}{2!} - \frac{\alpha^3 \bar{m}_3}{3!} + \dots \right]$$

---

When  $y$  is negative,  $\log y$  is, of course, not real.

When  $1-a > 0$  and is rational, so that  $y^{1-a}$  can be written as  $y^{p/q}$ , where  $p$  and  $q$  are integers, then if  $q$  is even and  $p$  odd,  $U(y)$  would be imaginary when  $y$  is negative. If  $q$  is odd but  $p$  is even, the utility of negative wealth would be positive and the marginal utility of wealth would be negative. When both  $p$  and  $q$  are odd, the utility for negative wealth would indeed be negative and the marginal utility positive, but then the marginal utility would increase with wealth, when  $y$  is negative.

When  $1-a$  is irrational,  $y^{1-a}$  is undefined when  $y$  is negative.

which must also converge, provided that the distribution of  $h$  has finite moments.<sup>5</sup> If higher order moments are of the order of magnitude of the corresponding powers of  $S$  (as, for instance, in the case of a normal distribution where

$$\bar{m}_{2k} = \frac{(2k)!}{2^k k!} S^{2k}$$

the series would converge rapidly if  $\alpha S$  is a

<sup>5</sup> Lately Benoit Mandelbrot and Eugene Fama (1965, 1971) claim that returns on speculative holdings of commodities or common stocks appear to be distributed according to nonnormal members of stable Paretian distributions with no finite moments. Their claims, however, are open to at least two grave doubts. First, even if it is granted that percentage daily fluctuations of stock or commodity prices do appear to conform well with some nonnormal stable Paretian distribution within the very limited range of observed daily changes, there is no guarantee that fluctuations of stock and commodity prices must conform with such a distribution over its entire range from  $-\infty$  to  $+\infty$ . Indeed, to make such a fantastic claim would clearly run afoul of the obvious constraint that the decline in stock or commodity prices cannot possibly exceed 100 percent. What Mandelbrot and Fama have in mind must be some sort of *truncated* stable Paretian distributions, which, being truncated, would not have infinite moments, nor would be stable under addition.

Secondly, it does not seem to me possible to find an acceptable utility function (in the sense of one that satisfies the above four conditions) that can be applied to a stable Paretian distribution with no finite moments and yet yields a finite positive expected utility—a necessary condition that investments with returns of such probability distributions would be desired at all. For instance, the negative exponential utility function, which is the only one that I can think of that has the required properties for a risk averter over the full range of wealth variation from  $-\infty$  to  $+\infty$ , would, when applied to a Cauchy distribution (the only nonnormal stable Paretian distribution whose analytical function is known), invariably yields an expected utility of  $-\infty$ , so long as the mean (or rather median) of the distribution is finite. That is

$$E(U) = B \int_{-\infty}^{\infty} \left\{ 1 - \frac{e^{-y}}{\pi[1 + (y-m)^2]} \right\} dy = -\infty$$

for any finite value of  $m$ .

So far Mandelbrot and Fama have failed to show what kind of utility functions investors must have for them to put a finite positive valuation on any prospect of return with an infinite variance, and whether the indicated utility functions would have the necessary properties for an ordinary risk-averse investor.

small fraction. The crucial question is, therefore, whether we can reasonably expect that this is the case for most problems of portfolio analysis even though  $\alpha$  is here unspecified in magnitude.

I think we can, if we stipulate the plausible speed with which the ultimate bliss, i.e., the upper limit of utility represented by  $B$ , is to be approached as wealth increases. That is, if we make the reasonable assumption that a tenfold increase of the total wealth of a normal individual is still unlikely to bring him to within, say, 1 percent (or 0.1 percent) of his ultimate bliss, then we may write

$$B(1 - e^{-10\alpha\bar{y}}) \leq B(1 - 0.01), \text{ or}$$

$$B(1 - 0.001),$$

$$\therefore \alpha \leq \frac{\ln 100}{10\bar{y}} = \frac{0.46}{\bar{y}},$$

$$\left( \text{or } \frac{\ln 1000}{10\bar{y}} = \frac{0.69}{\bar{y}} \right)$$

Anyway, it seems safe to maintain that  $\alpha = k/\bar{y}$ , where  $0 < k < 1$ . Thus we see that the condition that  $\alpha S < 1$  is equivalent to  $S < \bar{y}/k$ , or that risk be smaller than, say, twice his total wealth, including his human capital (i.e., the capitalized value of his earning capacity) as well as all his financial and physical assets. When we are discussing the behavior of a rational portfolio investor, this condition would undoubtedly hold, for it is extremely unlikely that he would voluntarily assume such enormous risk as would easily make himself bankrupt.

Indeed, if  $\alpha S = kS/\bar{y}$  is normally fairly small, then for a fair approximation in many problems, we may safely neglect terms with higher moments than the second or the third, even though, in absolute magnitude,  $S$ , i.e., risk, is not infinitesimal. For if we adopt the customary "pure number" measures for skewness and peakedness (kurtosis), viz.,  $\mu_3 = \bar{m}_3/S^3$  and

$\mu_4 = \bar{m}_4/S^4$ , respectively, equation (4) may be rewritten as

$$(4') \quad E[U(y)] = \beta - \beta_e^{-\alpha y} \left[ 1 + \frac{\alpha^2 S^2}{2!} - \frac{\alpha^3 S^3}{3!} \mu_3 + \frac{\alpha^4 S^4}{4!} \mu_4 - \dots \right]$$

Thus if, for instance, the risk under our consideration (as measured by the standard deviation) ranges only from zero up to, say, 10 percent of the individual's expected value of total wealth, then, assuming  $k=1/2$  roughly,  $\alpha S \leq 1/20$ . The coefficient for the pure number measure of skewness in the brackets would be absolutely less than  $1/48,000$ ; and that for peakedness would be less than  $1/3,840,000$ . It would seem, therefore, that a fair approximation of the expected utility for most practical purposes can be obtained by considering the mean and the variance only. As the ratio of risk,  $S$ , to the mean value of total wealth increases, however, the mean-variance analysis would become less and less accurate, and higher order central moments, in particular  $\bar{m}_3$ , would have to be taken into consideration.

Equation (4) clearly shows that the influence of skewness  $\bar{m}_3/S^3$  on the expected utility is positive. That is to say, a positive skewness of the distribution is a desirable feature, and, other things being equal, a greater skewness would increase the expected utility.<sup>6</sup> This is not a result peculiar to the assumption of a negative exponential utility function, but may be shown to be a general pattern of behavior towards

uncertainty on the part of all risk-averse individuals with decreasing or constant absolute risk-aversion with respect to increases in wealth. From equation (2), we may see that the coefficient of  $\bar{m}_3$  is  $U'''/3!$ . For a person with decreasing or constant absolute risk-aversion,  $U'''$  must be positive; for absolute risk-aversion is defined as  $-U''/U'$ , and

$$\frac{d}{dy} \left[ \frac{-U''}{U'} \right] = \frac{-U'U''' + (U'')^2}{(U')^2} \leq 0,$$

only if  $U''' \geq (U'')^2/U' > 0$ . Thus if we regard the phenomenon of increasing absolute risk aversion as absurd, we must acknowledge that a normal risk-averse individual would have a preference for skewness, in addition to an aversion to dispersion (variance) of the probability distribution of returns.<sup>7</sup>

It is interesting to note that Harry Markowitz, another pioneer of the  $E$ - $S$  analysis, once remarked that

... the third moment of the probability distribution of returns from the portfolio may be connected with a propensity to gamble. For example, if the investor maximize utility ( $U$ ) which depends on  $E$  and  $V$  ( $U = U(E, V)$ ,  $\partial U/\partial E > 0$ ,  $\partial U/\partial V < 0$ ), he will never accept an actuarially fair bet. But if  $U = U(E, V, M_3)$  and if  $\partial U/\partial M_3 \neq 0$ , then there are some fair bets which would be accepted. [pp. 90-91]

Nevertheless, as we have shown above, skewness preference ( $\partial U/\partial M_3 > 0$ ) is certainly not necessarily a mark of an inveterate gambler, but a common trait of a risk-averse person with decreasing or constant absolute risk-aversion. I cannot, therefore, go along with Markowitz in taking the view that since gambling is to be avoided,

<sup>6</sup> The connection between decreasing absolute risk aversion and  $U'''$  being positive was also pointed out by Joseph Stiglitz (p. 279). But he did not point out that this implies a skewness preference, since he did not operate with moments at all. A note of caution must be entered here; viz., that  $\mu_3 = \bar{m}_3/S^3$  is not a perfect measure of skewness. Although a symmetric distribution would necessarily have a zero  $\bar{m}_3$  or  $\mu_3$ , a zero  $\bar{m}_3$  or  $\mu_3$  does not imply that the distribution is symmetrical. Nevertheless,  $\mu_3$  is usually regarded by applied statisticians as a highly satisfactory measure of skewness.

<sup>7</sup> Samuelson in his latest article (1970) has also pointed out that the introduction of the third and higher moments would improve the accuracy of the mean-variance analysis, but he did not say whether the third central moment should have a positive or negative influence on the expected utility.

the third moment needed not be considered in portfolio analysis.

Anyway, skewness preference must be a fairly prevalent pattern of investor's behavior, for modern financial institutions provide a number of devices for investors to increase the positive skewness of the returns of their investments: for example, the organization of limited liability joint stock companies, prearranged stop-loss sales on the stock and commodity markets, puts and calls in stocks, etc., which otherwise would perhaps not have been developed.

If we plug in a constant elasticity utility function into (1) and (2), we would reach more or less the same conclusions. In this case, however, we must keep in mind the limitations of constant elasticity utility functions, viz., that they may be undefined, not real, or otherwise inapplicable as a utility function for zero or negative wealth and that their expansions will not converge unless the deviations from the mean value of wealth remain smaller than the mean itself. That is

$$\begin{aligned}
 U(y) &= \frac{1}{1-a} (\bar{y} + h)^{1-a} \quad (a > 0) \\
 (5) \quad &= \frac{1}{1-a} (\bar{y})^{1-a} + \frac{h}{(\bar{y})^a} - \frac{a}{2!} \frac{h^2}{(\bar{y})^{1+a}} \\
 &\quad + \frac{a(1+a)}{3!} \frac{h^3}{(\bar{y})^{2+a}} - \dots,
 \end{aligned}$$

or, with the logarithm form of constant elasticity utility function,

$$\begin{aligned}
 U(y) &= \log y = \log \bar{y} \left( 1 + \frac{h}{\bar{y}} \right) \\
 (5') \quad &= \log \bar{y} + \frac{h}{\bar{y}} - \frac{1}{2} \frac{h^2}{(\bar{y})^2} + \frac{1}{3} \frac{h^3}{(\bar{y})^3} \\
 &\quad - \dots,
 \end{aligned}$$

will not converge unless  $|h| < \bar{y}$ . As long as this constraint is satisfied, then

$$\begin{aligned}
 E[U(y)] &= \frac{1}{1-a} (\bar{y})^{1-a} - \frac{a}{2} \frac{S^2}{(\bar{y})^{1+a}} \\
 (6) \quad &\quad + \frac{a(1+a)}{3!} \frac{\bar{m}_3}{(\bar{y})^{2+a}} - \dots
 \end{aligned}$$

or, with a *log* utility function as in (5'),

$$\begin{aligned}
 (6') \quad E[U(y)] &= \log \bar{y} - \frac{1}{2} \frac{S^2}{(\bar{y})^2} + \frac{1}{3} \frac{\bar{m}_3}{(\bar{y})^3} \\
 &\quad - \dots,
 \end{aligned}$$

must also be a convergent series.<sup>8</sup>

In this case, the convergence of the series is slower than in the case of negative exponential utility. Nevertheless, if the risk under consideration remains smaller than, say, 10 percent of the expected value of total wealth, the effects of higher moments might still be fairly safely neglected. For instance, (6') can be rewritten as

$$\begin{aligned}
 (6'') \quad E[U(y)] &= \log \bar{y} - \frac{1}{2} \frac{S^2}{(\bar{y})^2} + \frac{1}{3} \frac{S^3}{(\bar{y})^3} \mu_3 \\
 &\quad - \frac{1}{4} \frac{S^4}{(\bar{y})^4} \mu_4 + \dots
 \end{aligned}$$

where, as before,  $\mu_3 = \bar{m}_3/S^3$  and  $\mu_4 = \bar{m}_4/S^4$ . If  $S/\bar{y} < 1/10$ , then the coefficient of the pure number measure of skewness is less than  $1/3000$ , and that of peakedness is absolutely less than  $1/40000$ .

Again we see that a person with a constant elasticity utility function (hence, risk-averse and with decreasing absolute risk-aversion) must have skewness preference, as the sign of the term with  $\bar{m}_3$  in (6) or (6') is necessarily positive.

Thus we see that the *E-S* analysis can be justified as a useful approximate method for portfolio selection even when utility function is not quadratic nor are the distributions of investment outcomes always normal. The only condition re-

<sup>8</sup> Equations (6) and (6') are really quite similar, especially since Arrow argues that "the relative risk aversion" (which in the case of (6) is  $yU''/U' = a$ ) "must hover around 1" (p. 37).

quired is that the risk  $S$  assumed by the investor must remain a fairly small fraction of his total wealth, including not only his entire net worth but also his human capital.

The necessity to include human capital as well as nonhuman net worth should be emphasized here; for as long as we recognize that utility function of wealth is non-linear, and that marginal utility of wealth diminishes with the increase in total wealth, we cannot apply a separate utility (or welfare) function to each separate investment or item of wealth. Since an individual's own personal earning capacity is as much a source of income as his physical and financial assets, his human capital must be included in his utility (or welfare) function with all other items of wealth. Thus the relevant  $S/\bar{y}$  ratio in the Taylor's expansion of the expected utility function should not be the ratio of the standard deviation of the returns of risk investment (or portfolio) to the mean of those returns alone, but should always be the ratio of total risk including the former to the mean value of the total wealth of the person concerned. For a normal portfolio investor anyway, this ratio probably can be expected to be a fairly small fraction, and hence the  $E$ - $S$  analysis might be a fair approximation with much wider application than its critics are willing to concede.

It might still be objected: if the  $E$ - $S$  analysis is only a method of approximation applicable to cases where risk is a small fraction of total wealth, why bother with it at all and not go directly to the general principle of maximizing expected utility. This is indeed the approach favored by many economists and some important works on economic behavior under uncertainty have been achieved without the use of  $E$ - $S$  approach or any moments.<sup>9</sup> This

approach is quite adequate for theoretical economists who are interested only in the direction of change induced by some hypothetical shift in parameters. For practical investors, who desire a fairly accurate quantitative guidance for action, however, it is not of much help. Since acceptable utility functions are non-linear and non-polynomial, they are generally rather difficult to integrate when multiplied by the density function of random returns of wealth.<sup>10</sup> Generally, it would be more convenient if we should be able to expand the utility function into a Taylor's series and then integrate it term by term. As we have shown above, this approach leads directly to the approximation of the expected utility function by a function of the moments of the distribution, and if convergence warrants it, we might use the terms with the mean and the variance alone for a fair approximation. Furthermore, ordinarily the investor does not know what the exact shapes of the distribution functions of investment returns are. Usually, as in so many problems of applied statistics, one has only some estimates of the locations, dispersions and perhaps some vague idea of the degrees of skewness of the distributions to work on. These facts constitute the basic rationale of the mean-variance analysis.

In a broad sense, the  $E$ - $S$  analysis versus the general principle of maximizing expected utility is analogous to the consumers' surplus analysis versus the general equilibrium analysis of welfare gains. For small risk relative to total wealth, the  $E$ - $S$  analysis is adequate enough and can be handled quantitatively with much

<sup>9</sup> Stiglitz's contribution on taxation and risk-taking, and David Levhari and T. N. Srivasan and L. J. Mirman on uncertainty and savings are good examples of this approach.

<sup>10</sup> The case of a  $\log$  utility function combined with a lognormal distribution for returns of investment seems to be an exception. See Feldstein's 1969 paper which will be discussed below. However, the peculiar ease of integrating the expected utility function would disappear when one realizes that there might be a certain core in the investor's total wealth, and that the lognormal distribution might apply only to the marginal investment on some risky assets.

greater facility. Indeed, if the risk is so small that, over the range of random variations of wealth, changes in the marginal utility of wealth (money) can be neglected (which is the equivalent of Marshall's justification for the measurement of consumers' surplus by the triangle under the demand curve), then the increment in expected utility of wealth may be approximated by the utility of the increment of expected wealth. There would be no need to consider even the variance. Surely, as Tobin (1969) claimed in its defense, the  $E$ - $S$  analysis represents a major step forward for cases, where, over the range of variations of wealth, decrease or increase in the marginal utility of wealth cannot be neglected.

## II. Borch's and Feldstein's Paradoxes and the Impossibility that Slopes of $E$ - $S$ Indifference Curves Equal or Exceed $45^\circ$

The above discussion would enable us to give a satisfactory explanation of the paradoxes posed by Borch and Feldstein's criticisms of the  $E$ - $S$  analysis.

Borch's criticism, as mentioned in the beginning, is that any system of upward sloping indifference curves in the  $E$ - $S$  plane can be shown to be inconsistent with the basic axiom of choice under uncertainty. In Figure 1,  $(E_1, S_1)$  and  $(E_2, S_2)$  are any two points on a given  $E$ - $S$  indifference curve. Borch shows that these two vectors in  $E$  and  $S$  may be represented by two Bernoulli distributions  $(x, p, y_1)$  and  $(x, p, y_2)$ , where

$$(7) \quad x = \frac{S_1 E_2 - S_2 E_1}{S_1 - S_2}$$

$$(8) \quad p = \frac{(E_2 - E_1)^2}{(E_2 - E_1)^2 + (S_2 - S_1)^2}$$

$$(9) \quad y_1 = E_1 + S_1 \frac{S_2 - S_1}{E_2 - E_1}$$

and

$$(10) \quad y_2 = E_2 + S_2 \frac{S_2 - S_1}{E_2 - E_1}$$

as it may be easily verified that the mean and standard deviation of  $(x, p, y_i)$  would be  $E_i$  and  $S_i$ . Since  $E_2 > E_1$ , and  $S_2 > S_1$ ,  $y_2$  must be  $> y_1$ . Thus  $(x, p, y_2)$  is axiomatically preferable to  $(x, p, y_1)$ ; yet they are supposed to be on the same indifference curve.

This apparent contradiction can be explained as follows. Let us first assume that the relevant section of the indifference curve nowhere has slope equal to or greater than unity (or  $45^\circ$ ). The  $E$ - $S$  analysis, as we have shown above, is based on the approximation of the expected utility function by its expansion involving no terms higher than the second order. If, however, the third central moment (skewness) of the distribution is known to be positive, its omission would imply a downward bias in the expected utility, as  $U'''$  must be positive for a risk-averter (one with upward rising  $E$ - $S$  indifference curves), who does not have an increasing absolute risk-aversion (which is generally considered as absurd).

Now on the section of an indifference curve with slope everywhere less than 1, the two Bernoulli distributions, constructed in the above manner to represent any two arbitrary points on the curve, must be positively skewed; and the one representing the point to the northeast of the other must have a greater skewness. For

$$(11) \quad P = \frac{(E_2 - E_1)^2}{(E_2 - E_1)^2 + (S_2 - S_1)^2} < \frac{1}{2},$$

when  $(E_2 - E_1) < (S_2 - S_1)$ , as is implied by our assumption that the slope of the relevant section of the  $E$ - $S$  indifference curve is nowhere equal to, or greater than, 1. And  $\bar{m}_3$ , a measure for skewness, is defined as

$$\begin{aligned}
 (12) \quad \bar{m}_3 &= [x - py_i - (1 - p)x]^3(1 - p) \\
 &\quad + [y_i - py_i - (1 - p)x]^3p \\
 &= p(1 - p)(1 - 2p)(y_i - x)^3 \geq 0, \\
 &\quad \text{as } \begin{cases} 0 < p < \frac{1}{2} \\ p = \frac{1}{2} \\ 1 > p > \frac{1}{2} \end{cases}
 \end{aligned}$$

since it may easily be shown that  $y_i$  is always  $> x$ .<sup>11</sup> The cases where  $p=0$  or 1, being trivial, are excluded. Given that  $p < 1/2$  so that both distributions would be positively skewed, the distribution with the probability  $p$  of gaining  $y_2$  (represented by the point  $(E_2, S_2)$ ) must be more positively skewed than the distribution with the probability  $p$  of gaining  $y_1$  (represented by the point  $(E_1, S_1)$ ), as  $y_2$  is  $> y_1$ .

Thus the contradiction between the implication of the indifference curves and the fact that  $(x, p, y_2)$  must be preferred to  $(x, p, y_1)$  can be explained by the fact that, in the  $E$ - $S$  analysis, the advantage of the greater skewness of  $(x, p, y_2)$  as compared with that of  $(x, p, y_1)$  is neglected on the assumption that the influence of skewness is of a smaller order of magnitude than that of variance and mean. So long as this assumption about the order of magnitude is justifiable on the basis of the relative smallness of the standard deviation to total wealth, the treatment of  $(x, p, y_1)$  and  $(x, p, y_2)$  as approximately indifferent is also justifiable.

This explanation of the paradox, of course, would not be valid if the slope of  $E$ - $S$  indifference curves should rise to 1 or greater. For if so, we can always pick two points  $(E_1, S_1)$  and  $(E_2, S_2)$  on a given in-

difference curve such that  $E_2 - E_1$  exactly equals  $S_2 - S_1$ , in which case

$$p = \frac{(E_2 - E_1)^2}{(E_2 - E_1)^2 + (S_2 - S_1)^2} = \frac{1}{2},$$

and hence the two Bernoulli distributions constructed in the manner suggested by Borch would both have zero skewness. Or we may even pick the two points in such way that  $(E_2 - E_1) > (S_2 - S_1)$ , in which case,  $p > 1/2$ , and, hence, the two Bernoulli distributions  $(x, p, y_1)$  and  $(x, p, y_2)$  would both be negatively skewed, and the negative skewness would be absolutely larger in the case of  $(x, p, y_2)$ ,  $y_2$  being  $> y_1$ , as is obvious from equation (12).

In these cases, we cannot use the neglected influence of skewness to explain why  $(x, p, y_2)$  and  $(x, p, y_1)$  may be represented by two points  $(E_2, S_2)$  and  $(E_1, S_1)$  on the same indifference curve, yet, strictly speaking, the former must be preferred to the latter. However, this is not to be regarded as an indication of the inherent inconsistency of  $E$ - $S$  indifference map as an approximate representation of the scale of preference for returns and risk. Rather it should be interpreted as an indication that such indifference curves should never be drawn with slope going up to  $45^\circ$  or even greater. This is an important fact which users of  $E$ - $S$  indifference curves usually overlook, sometimes with rather absurd results.<sup>12</sup>

<sup>12</sup> For instance, G. O. Bierwag and Myron Grove by neglecting the constraint on the slopes of  $E$ - $S$  indifference curves, were led to the strange conclusion that the indifference curves for any two risky assets (the yield-risk ratios of which are unspecified) are closed curves, which are drawn in their Figure 3 as concentric circles, with the center at some bliss point, apparently not related to the concept of an ultimate bliss in certain types of utility functions such as the negative exponential function. This would imply that a portfolio of, say, 10 shares each of, say, GM and IBM might yield the same expected utility as a portfolio of, say, a million shares each of these two stocks; and that, in the latter situation, the owner could increase his expected utility by destroying some of both assets—a phenomenon which

<sup>11</sup> To show that  $(y_i - x) > 0$ , it is necessary only to show that  $y_1 > x$ . From (7) and (9), we may obtain

$$\begin{aligned}
 y_1 - x &= \frac{E_1(E_2 - E_1) + S_1(S_2 - S_1)}{(E_2 - E_1)} - \frac{S_2E_1 - S_1E_2}{(S_2 - S_1)} \\
 &= \frac{S_1(E_2 - E_1)^2 + S_1(S_2 - S_1)^2}{(E_2 - E_1)(S_2 - S_1)} > 0
 \end{aligned}$$

In fact, if we go back to the foundation of the *E-S* analysis, viz., the justifiability of approximating the expected utility function with its quadratic expansion, i.e., equation (2') above,

$$E[U(y)] = U(\bar{y}) + U'' \frac{S^2}{2}$$

we could readily see that the slope of the *E-S* indifference curves,

$$\left. \frac{dy}{ds} \right| = \frac{-SU''}{U'}$$

$$(E[U] = \text{constant})$$

could not be equal to or greater than 1; for

$$\frac{-SU''}{U'} \geq 1$$

would imply that

$$U'' + SU'' \leq 0$$

which means that, provided the quadratic approximation is a close approximation within this range, the marginal utility of wealth would be brought to zero or to a negative value by a mere deviation of one standard deviation of the actual value of wealth from its mean.

To use the negative exponential utility

Bierwag and Grove call "contamination by risk," but which is obviously contrary to our common sense.

Now if we realize that the *E-S* indifference curves would never have slopes greater than unity, then a ray from the origin in the positive *E-S* space with a slope greater than unity would never cut the same *E-S* indifference curve twice. Mapped into the two asset space, it means that so long as none of the two assets has a risk greater than its own expected terminal value (which can certainly be said for either GM or IBM stocks, even though they are not riskless), no rays from the origin representing proportionate variations of given portfolios of these two stocks should cut any asset indifference curves more than once. Thus the paradoxical conclusion of Bierwag and Grove can be shown to be incorrect, because they assumed that *E-S* indifference curves can slope up to be asymptotic to the vertical, so that any rays, with slope of 45° or more, can still cut each *E-S* indifference curve at two points. Actually the so-called phenomenon of contamination by risk cannot take place unless the risk of the portfolio involved is greater than its expected terminal value.

function as an illustration, if the expected utility of wealth is sufficiently closely approximated by

$$E[U(y)] = B - Be^{-\alpha \bar{y}} \left[ 1 + \frac{\alpha^2}{2} S^2 \right],$$

then

$$(13) \quad \left. \frac{d\bar{y}}{dS} \right| = \alpha S / \left( 1 + \frac{\alpha^2 S^2}{2} \right) \\ (E[U] = \text{constant})$$

which is necessarily  $< 1$ . Indeed, it must also be smaller than  $\alpha S$ , which must itself be considerably smaller than 1; for otherwise the use of the *E-S* analysis could not be justified.

Similar conclusions can be obtained with a constant elasticity utility function of either the Cobb-Douglas form or the logarithmic form, so long as the expected utility of wealth can be sufficiently closely approximated by their quadratic expansions, i.e.,

$$E[U(y)] = \frac{1}{1-a} (\bar{y})^{1-a} - \frac{a}{2} \frac{S^2}{(\bar{y})^{1+a}}$$

or

$$E[U(y)] = \log \bar{y} - \frac{1}{2} \frac{S^2}{(\bar{y})^2}$$

For then the slope the *E-S* indifference curves would be

$$(14) \quad \frac{d\bar{y}}{dS} = \frac{\alpha S}{\bar{y}} / \left[ 1 + \frac{a(1+a)}{2} \frac{S^2}{(\bar{y})^2} \right],$$

or

$$(14') \quad \frac{d\bar{y}}{dS} = \frac{S}{\bar{y}} / \left[ 1 + \frac{S^2}{(\bar{y})^2} \right],$$

respectively. In both cases,  $d\bar{y}/dS$  must be smaller than 1. It must also be smaller than  $\alpha S/\bar{y}$  or  $S/\bar{y}$ , as the case may be, which must themselves be considerably smaller than 1 for the use of *E-S* analysis to be justified.

Even in the cases of quadratic utility function and of normally distributed investment outcomes, where the  $E$ - $S$  analysis is considered applicable without restriction, we shall also see that the slope of  $E$ - $S$  indifference curves would be smaller than 1 in their relevant range. With a quadratic utility function, the impossibility of the derived  $E$ - $S$  indifference curves to be steeper than  $45^\circ$  has already been pointed out by Tobin (1958).

In the case where all investment returns are assumed to be normally distributed, Tobin (1958, p. 75) has shown that for each level of  $E[U]$ ,

$$(15) \quad \frac{d\bar{y}}{dS} = \frac{\int_{-\infty}^{\infty} z U'(\bar{y} + Sz) \phi(z; 0, 1) dz}{\int_{-\infty}^{\infty} U'(\bar{y} + Sz) \phi(z; 0, 1) dz},$$

where  $\phi(z; 0, 1)$  is the normal density function with zero mean and unity variance. It is not possible to evaluate this expression quantitatively unless the utility function is specified, although it may be demonstrated that for a risk-averter, the indifference curves would be upward sloping and convex downward. It is clear that not every utility function can be employed with untruncated normally distributed returns, as the range of variations reaches from  $-\infty$  to  $\infty$ . The quadratic utility function and the constant elasticity utility function are clearly ruled out.<sup>13</sup> With a negative exponential utility

function  $U = B(1 - e^{-\alpha y})$ , however, (15) can be evaluated by a different approach; for with this utility function, the expected utility may be expanded into a series of deviations of  $y$  from its mean, as shown in (4).

Furthermore, with normal distribution, all odd order central moments vanish, and all even order higher central moments can be expressed as functions of  $S^2$ , viz.,<sup>14</sup>

$$\bar{m}_{2k} = \frac{(2k)!}{2^k k!} S^{2k}$$

Hence

$$(16) \quad \frac{d\bar{m}_{2k}}{dS} = \frac{(2k)!}{2^{k-1}(k-1)!} S^{2k-1}$$

Differentiating (4) with respect to  $\bar{y}$  and  $S$  after eliminating all terms with odd order moments and substituting (16) into the result and then setting both sides to zero, we obtain equation (17). Comparing (17) with (13) above, we can see that the slope of indifference curves computed with the quadratic approximation is a very good approximation indeed, when  $\alpha S$  is a smaller fraction.

As we have shown above, the parameter  $\alpha$  of a negative exponential utility function must be understood to be of the magnitude of  $k/\bar{y}$ , where  $0 < k < 1$ . Thus even though we are not concerned here with the rapidity of convergence, as we were

<sup>13</sup> According to Samuelson (1967, p. 9), an investor with a  $\log$  utility function would evaluate any normally distributed investment outcome as having an expected

utility of  $-\infty$ . The same can be said of an investor with a constant elasticity utility function with an elasticity greater than unity, since in this case, as in the case of  $\log$  utility function, utility would approach minus infinity as wealth approaches zero from above.

<sup>14</sup> See, e.g., B. W. Lindgren, pp. 87 and 89.

$$(17) \quad \frac{d\bar{y}}{dS} = \frac{\alpha S \left( 1 + \frac{\alpha^2}{2} S^2 + \frac{\alpha^4}{8} S^4 + \dots + \frac{\alpha^{2k}}{2^k k!} S^{2k} + \dots \right)}{\left( 1 + \frac{\alpha^2}{2} S^2 + \frac{\alpha^4}{8} S^4 + \dots + \frac{\alpha^{2k}}{2^k k!} S^{2k} + \dots \right)} = \alpha S$$

when using quadratic approximation, still for all practical purposes in portfolio analysis,  $\alpha S$  should generally be treated as smaller than 1. For  $\alpha S$  to be greater than 1 would imply that the investor concerned is assuming a risk (i.e., the standard deviation of the value of his wealth) much larger than the expected value of his total wealth, including his human capital as well as all other nonhuman assets. This must be regarded as a most unlikely situation for a risk-averse investor, and any person with this kind of risk-asset ratio would certainly be considered extremely uncreditworthy. Thus even in this case, slopes of  $E$ - $S$  indifference curves should be understood to be normally much smaller than 1 in the range relevant for portfolio investors with risk-aversion, although it is not theoretically impossible for their slope to exceed 1.

It is to be noted, however, that in this case, the indifference curves thus traced out would incorporate all the indirect influences upon the expected utility of  $S$  through its linkages with all the even order higher moments, that are peculiar to normal distributions only.<sup>15</sup> Such a system of  $E$ - $S$  indifference curves, therefore, cannot be applied to cases where investment outcomes might have any other types of distributions. Certainly, it would be quite senseless to fit, say, two Bornoulli distributions to two points on any indifference curve of this system to try to prove its logical inconsistency.

The paradox posed by Feldstein's criticism can be clarified in the similar manner. He has shown that, with a  $\log$  utility function and a lognormal distribution for investment outcomes, the  $E$ - $S$  indifference curves for a risk-avertter need not always

be convex downwards but would change from convex to concave, once the standard deviation  $S$  exceeds a certain crucial proportion of the mean of total wealth, viz.,  $\bar{y}/\sqrt{2}$ . That is, if the expected utility of wealth is

$$\begin{aligned} E[U(y)] &= \int_{-\infty}^{\infty} \log y f(y) dy \\ (18) \quad &= \int_{-\infty}^{\infty} x \phi(x) dx = \mu \end{aligned}$$

where  $f(y)$  is a lognormal distribution and, hence, the logarithm of  $y$ , viz.,  $x$ , has a normal distribution  $\phi(x)$  with a mean  $\mu$  and a variance  $\sigma^2$ , then

$$(19) \quad E[U(y)] = \log \bar{y} - \frac{1}{2} \log \left( \frac{S^2}{\bar{y}^2} + 1 \right)$$

This follows readily from the well-known formulae for the mean and variance of a lognormal variable in terms of the mean  $\mu$  and variance  $\sigma^2$  of the logarithm of the variable, viz.,  $\bar{y} = E(y) = \exp(\mu + \sigma^2/2)$ , and  $S^2 = E(y - \bar{y})^2 = \bar{y}^2 (\exp \sigma^2 - 1)$ . Differentiating (19) with respect to  $\bar{y}$  and  $S$ , and setting both sides to zero we get

$$(20) \quad \frac{d\bar{y}}{dS} = \frac{S}{\bar{y}} \left/ \left( 1 + 2 \frac{S^2}{\bar{y}^2} \right) \right.$$

which is  $>0$ , but always  $<1$ . (See Feldstein, p. 8.)

Comparing this with equation (14') above, again we see that the slope of  $E$ - $S$  indifference curve, which we obtained there by using the quadratic approximation of the  $\log$  utility function without specifying the probability distribution of investment returns, is indeed a close approximation for this case, provided that  $S/\bar{y}$  is a small fraction.

Taking the second derivation along the indifference curve, it is found that

$$(21) \quad \frac{d^2\bar{y}}{dS^2} = \frac{\bar{y}(\bar{y}^2 - 2S^2)(\bar{y}^2 + S^2)}{(2S^2 + \bar{y}^2)^3},$$

<sup>15</sup> As pointed out by Samuelson (1967, p. 11), such indifference contours are not drawn from knowledge of the decision maker's risk preferences alone. They must, therefore, be redrawn for each new probability distribution of random outcomes.

which is positive when  $S^2/\bar{y}^2 < 1/2$ , but becomes negative when  $S^2/\bar{y}^2 > 1/2$ , or  $S/\bar{y} > 0.707$ . (See Feldstein, p. 8.) The indifference curves would look something like the one in Figure 2. On the face of it, it appears that risk-aversion might eventually decrease as risk itself is increased!<sup>16</sup>

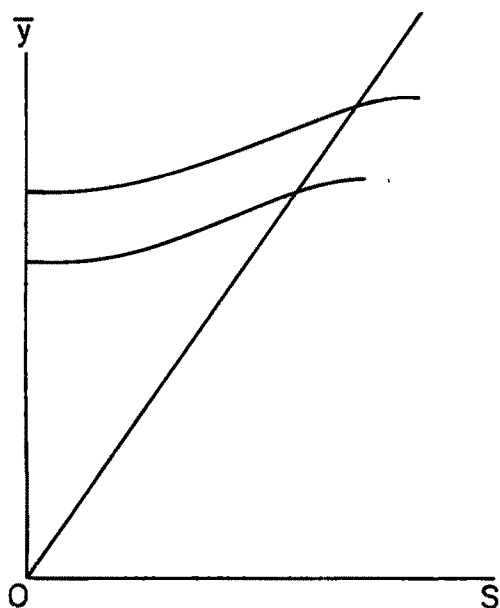


FIGURE 2

Such a hasty interpretation, of course, would be quite unwarranted. The reversal of the change in the curvature of  $E$ - $S$  indifference curves in this case is essentially due to the particular linkages between the mean and variance and higher moments of lognormal distributions. Although lognormal distributions have only two pa-

rameters, they are not symmetric but positively skewed. The third central moment of such a distribution  $f(y)$ , where  $\log y$  is normally distributed with a mean  $\mu$  and a variance  $\sigma^2$ , is given by the formula:<sup>17</sup>

$$(22) \quad \bar{m}_3 = \bar{y}^3(e^{3\sigma^2} - 3e^{\sigma^2} + 2)$$

Thus,  $\bar{m}_3$  is an increasing function of both  $\bar{y}$ , the mean, and  $S^2$ , the variance, of  $y$ . Holding  $\bar{y}$  constant, while we increase  $S$ , we find<sup>18</sup>

$$(23) \quad \frac{\partial \bar{m}_3}{\partial S} = 6\bar{y}S(e^{2\sigma^2} - 1) > 0$$

Since a person with a  $\log$  utility function must have a positive skewness preference, i.e.,  $\partial E[U]/\partial \bar{m}_3 = 1/3\bar{y}^3 > 0$ , and since in Feldstein's derivation of the slope of indifference curves the indirect influences of a change in variance through its linkages with the third and other moments are all taken into account, it is not surprising that the positive influence of the increasing skewness should partly offset the negative influence of increasing dispersion itself.

It is interesting to note that in this case also,  $E$ - $S$  indifference curves can never reach a slope of 1; for in this case, indifference curves would attain their maximum slope at their points of inflexion, i.e., at  $S/\bar{y} = \sqrt{0.5}$ . At the inflexion, the maximum slope is  $\max(d\bar{y}/dS) = (\sqrt{0.5}/2) = 0.3536$ , which is considerably smaller than 1.

It should also be noted that Feldstein's indifference curves are not, strictly speak-

<sup>16</sup> In fact, if we take the second derivatives of (13), (14) or (14') with respect to  $S$  along the respective indifference curves, we would find essentially similar results, viz., that in each case, the indifference curves would be convex downward only when  $S/\bar{y}$  is smaller than a certain value, but becomes concave when  $S/\bar{y}$  is greater than that value. However, since (13), (14), and (14') are derived from the quadratic approximation of the expected utility functions and such approximation is justified only when  $S/\bar{y}$  is very small, only the convex sections of the indifference curves would be meaningful.

<sup>17</sup> See e.g., Lindgren, p. 89, where the  $k$ th moment about zero of a lognormal distribution is given as  $E(y^k) = \exp(k\mu + (1/2)k^2\sigma^2)$ , and  $\text{Var}(y) = (\exp \sigma^2 - 1) \cdot (2\mu + \sigma^2)$ . From these, it may be worked out that

$$\bar{m}_3 = (\exp 3\sigma^2 - 3\exp \sigma^2 + 2)\bar{y}^3$$

<sup>18</sup> This means that we let  $\sigma^2$  increase, while reducing  $\mu$  to keep  $\bar{y} = \exp(\mu + \sigma^2/2)$  constant. Since  $S^2 = \bar{y}^2(\exp \sigma^2 - 1)$ ,

$$\frac{\partial \bar{m}_3}{\partial S} = \frac{\partial \bar{m}_3}{\partial \sigma^2} \cdot \frac{\partial \sigma^2}{\partial S} = 6\bar{y}S(\exp 2\sigma^2 - 1)$$

ing, the indifference curves for the expected value of wealth and its standard deviation per se, but for the mean and the standard deviation when they are known to be linked with all other higher moments in the particular ways inherent to the lognormal distribution; viz., through the relationship

$$(24) \quad \begin{aligned} E(y^k) &= \exp\left(k\mu + \frac{\sigma^2 k^2}{2}\right) \\ &= (\bar{y})^k \exp\left[(k-1) \frac{k\sigma^2}{2}\right] \end{aligned}$$

where  $\mu$  and  $\sigma^2$  are the mean and variance of  $\log y$ .<sup>19</sup> They are not derived from knowledge of the investor's utility function alone. They are, therefore, not applicable, unless the terminal value of the investor's total wealth is known to be distributed strictly according to the lognormal distribution.

In general, as Samuelson pointed out (1967), if we want to take into account in the indifference contours all the indirect effects of the mean and variance through their linkages with higher movements, then for a given utility function, we would have a different set of indifference curves for each probability distribution of the terminal value of the total wealth of the investor concerned. We can get a unique set of *E-S* indifference curves from a given utility function, only if we treat the truncated quadratic expansion of the expected utility function as an adequate approximation and neglect all higher moments. As we have seen, this is justifiable only when  $S/\bar{y}$  is fairly small. For small values of  $S/\bar{y}$ , Feldstein's indifference curves are certainly convex downwards and closely approximated by those derived with the quadratic approximation. It is only when  $S/\bar{y}$  is quite large, approaching  $\sqrt{0.5} = 0.7071$ , that problem of noncon-

vexity begins to appear. But before that happens, the mean-variance approximation would already have to be supplemented by consideration of the third and higher moments, or even abandoned altogether.

Thus it is clear that while the *E-S* analysis might be an adequate method of approximation for analyzing the investment behavior of small risk-takers (cautious portfolio investors who normally assume rather small risks relatively to their total wealth), for major risk-takers (entrepreneurs who regularly risk a major proportion of their total wealth), it would have to be supplemented or abandoned.

### III. Implications for the Theory of the Demand for Money

One important conclusion of the above discussion is that *E-S* indifference curves, insofar as their use is warranted, would never attain a slope of 45° or more. Usually, and for cautious portfolio investors in particular, the slopes of *E-S* indifference curves stay very much less than that. This conclusion has important implications for the theory of the demand for money.

In his pioneering work on the *E-S* analysis, "Liquidity Preference as Behavior Towards Risk," Tobin employed this analysis to explain why rational people would hold idle cash in their investment portfolios.<sup>20</sup> Although cash yields no income, the risk attached to other income-yielding assets may offset the attraction of their positive yields so that they are no more attractive to hold than cash at the margin of portfolio equilibrium. Diagrammatically, portfolio equilibrium is depicted as a tangency point (except in the

<sup>20</sup> The investment portfolio, or investment balance, is defined by Tobin as funds "that will survive all the expected seasonal excess of cumulative expenditures over cumulative receipts during the year ahead. They are balances which will not have to be turned into cash within the year" (1958, p. 66).

<sup>19</sup> See Lindgren, p. 89.

case of a "plunger") of a yield-risk opportunity curve with one of the  $E$ - $S$  indifference curves, which are usually drawn without much attention to the constraint on their slopes.

If it is realized that the slope of  $E$ - $S$  indifference curves usually stay well below  $45^\circ$ , this explanation for the demand for cash in the investment portfolio becomes highly implausible. For if there is any financial asset, the expected yield<sup>21</sup> of which is at least as large as the standard deviation of that yield, then that asset will surely be preferred to cash as investment. It may be shown with the help of Figure 3. Suppose the existing portfolio of

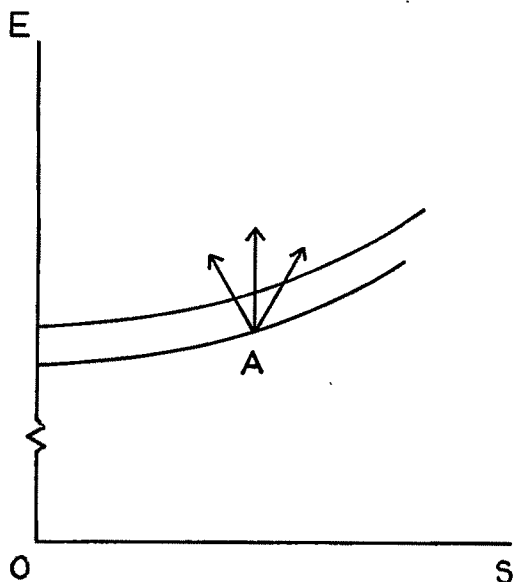


FIGURE 3

a given investor, the mean and risk of which are represented by the coordinates of the point  $A$  on a certain indifference curve, still contains some amount of idle cash holding. If there is any asset, the expected yield (as defined in the preceding

footnote) of which is no smaller than its own standard deviation, then by substituting this asset for cash, the investor can move upward from  $A$  on the  $E$ - $S$  plane at an angle with the horizontal axis at least equal to, but more likely greater than,  $45^\circ$ . See Figure 3. For by substituting that asset for cash, he increases the expected terminal value of his total wealth by the expected yield of that asset, whereas the increment in his total risk would generally be less than the risk of the newly acquired asset itself, so long as the yield of that asset is not perfectly correlated positively with the yield of the existing portfolio. Hence the angle of direction of the move is likely to be greater than  $45^\circ$  with respect to the horizontal, even if the expected yield of the new asset is no greater than its own standard deviation.<sup>22</sup>

<sup>22</sup> Suppose that the expected terminal value of an investor's total wealth at point  $A$  in Figure 3 is  $E(W) = W_0[1 + w_p u_p + (1 - w_p) u_c] = W_0(1 + w_p u_p)$ , where  $W_0$  is the initial value of his wealth,  $w_p < 1$  is the proportion of his wealth invested in some earning portfolio, the average expected rate of return of which is  $u_p$ , and  $u_c$  is the rate of return of cash holding, which is assumed to be zero. Since the terminal value of cash holding is supposed to be perfectly certain, the standard deviation of the terminal value of his wealth is simply  $S(W) = W_0 w_p \sigma_p$ , where  $\sigma_p$  is the standard deviation of the terminal value of his portfolio of earning assets.

Now suppose that he decides to transfer a small fraction of his wealth  $x$  from cash holding on to an asset  $i$  with an expected rate of return  $u_i$  and a variance  $\sigma_i^2$ , then the expected terminal value of his wealth would be  $E(W) = W_0[1 + w_p u_p + x u_i]$ , and its standard deviation would be

$$s(W) = W_0[w_p^2 \sigma_p^2 + x^2 \sigma_i^2 + 2w_p x \rho_{pi} \sigma_p \sigma_i]^{1/2},$$

where  $\rho_{pi} = \sigma_{pi} / \sigma_p \sigma_i$  is the correlation coefficient between  $u_p$  and  $u_i$ .

Differentiate both  $E(W)$  and  $S(W)$  with respect to  $x$ , we get  $\partial E / \partial x = W_0 u_i$ , and

$$\begin{aligned} \frac{\partial S}{\partial x} &= \frac{W_0(x \sigma_i^2 + w_p \rho_{pi} \sigma_p \sigma_i)}{(w_p^2 \sigma_p^2 + x^2 \sigma_i^2 + 2w_p x \rho_{pi} \sigma_p \sigma_i)^{1/2}} \\ &= W_0 \rho_{pi} \sigma_i, \quad (\text{at } x = 0) \end{aligned}$$

Therefore, by increasing  $x$  at  $A$ , where  $x=0$ , the  $(E, S)$  vector of the investor's total wealth can be moved in the direction of  $dE/dS = u_i / \rho_{pi} \sigma_i > u_i / \sigma_i$ , so long as  $\rho_{pi} < 1$ . If  $\rho_{pi} = 0$ , i.e., if  $u_i$  is totally uncorrelated with

<sup>21</sup> We define the expected yield of an asset as the expected terminal value (including accrued interest) of a dollar's worth of that asset at current price minus 1.

If the slope of the  $E$ - $S$  indifference curve at  $A$  is known to be smaller than  $45^\circ$ , then such a move would certainly carry him to a higher indifference curve. And so long as the slopes of high indifference curves remain smaller than  $45^\circ$ , such moves can be repeated, until the cash holding in the investment portfolio is exhausted.

Thus in order to demonstrate that investment or speculative cash balances really has no place at all in a rational investment portfolio, it is not necessary to show that there is some asset that "dominates" cash (i.e., as riskless as cash, but has in addition some positive yield). Nor is it necessary to show that there is some asset, the expected yield of which is larger than the maximum possible downward deviation from the expected yield. All that is necessary is to demonstrate that there is at least one asset, the expected yield of which is not smaller than its own standard deviation, or at least one asset with a positive expected yield, however small, which is uncorrelated, or negatively correlated, with the yield of the existing portfolio of the investor concerned. The existence of such assets would eliminate all demand for cash for the portfolio balance purpose.<sup>23</sup> Surely, there must be a host of

assets in modern financial markets, e.g., savings deposits and Treasury bills, that would satisfy these requirements.

Thus although the  $E$ - $S$  analysis was at first introduced by Tobin to explain liquidity preference in the sense of an investment demand for cash, in our defense of it against its critics, we actually find that it is quite incapable of doing what Tobin has expected of it. Rather it seems to indicate that there cannot be any investment demand for money for the so-called portfolio balance purpose. The demand for money must arise from the requirements of anticipated transactions and the precaution against contingencies calling for unplanned cash expenditures. Our discussion should, therefore, cast some strong doubts on the alleged superiority of the modern "wealth approach" or "portfolio balance approach" to monetary theory, which ironically seems to hold sway at present among both the neo-Keynesians and the neomonetarists.

Incidentally, the importance of skewness preference for major risk-takers should obviously be taken into consideration in problems of investment incentives. For instance, the effects of income tax on risk-taking should be examined not only with respect to its impacts on the mean and variance of investment returns after tax, but also with respect to its impact on the skewness of net returns. A progressive income tax or an income tax without adequate loss offset would certainly have greater adverse effect on the willingness to take risk than a proportional income tax with perfect loss offset that would leave the mean and variance after tax at the same levels.

$u_p$ , then  $dE/dS = \infty$ . In this case, the arrow drawn from  $A$  in Figure 3 would form an angle of  $90^\circ$  with the horizontal, i.e., it would point vertically upward. If  $u_i$  is negatively correlated with  $u_p$ , then  $dE/dS < 0$ . In this case, the arrow from  $A$  would form an angle greater than  $90^\circ$  with the horizontal, i.e., it would point north-west. In the latter two cases, there is no question but that higher indifference curves would be reached by moving along the arrow (i.e., by increasing  $x$  at  $A$ ).

<sup>23</sup> Actually Tobin has already noticed, incidentally to his discussion of the effect of a change in interest rate on liquidity preference, that, with a quadratic utility function, if the rate of interest of bonds  $r$  exceeds the risk of bonds  $\sigma_b$ , then tangency solution is impossible (i.e., the investor must hold all bonds no cash). To prove it, he used a rather devious method, suggested to him by Arthur Okun (See (1958) p. 79, especially fn. 1). Actually, the proof should be very simple. If  $r > \sigma_b$ , then the slope of the opportunity locus  $r/\sigma_b > 1$ . However, with a quadratic utility function, the slope of mean-standard deviation indifference curves must always be

smaller than 1. Hence, tangency solution is impossible. Now we have seen that this is true with all acceptable utility functions for a risk averter. Feldstein also noted that with a lognormal distribution of returns to bonds and a  $\log$  utility function, "the investor will be a 'plunger', holding only bonds, unless their variance is very high in relation to their expected yield" (p. 9).

## REFERENCES

- K. J. Arrow, *Aspects of the Theory of Risk-Bearing, Lectures*, Helsinki 1964.
- G. O. Bierwag, and M. A. Grove, "Indifference Curve in Asset Analysis," *Econ. J.*, June 1966, 76, 337-43.
- K. Borch, "A Note on Uncertainty and Indifference Curves," *Rev. Econ. Stud.*, Jan. 1969, 36, 1-4.
- P. H. Cootner, *The Random Character of Stock Market Prices*, Cambridge, Mass. 1964.
- E. F. Fama, "The Behavior of Stock Market Prices," *J. Bus. Univ. Chicago*, Jan. 1965, 38, 34-105.
- , "Risk, Return and Equilibrium," *J. Polit. Econ.*, Jan./Feb. 1971, 79, 30-35.
- M. S. Feldstein, "Mean-Variance Analysis in the Theory of Liquidity Preference and Portfolio Selection," *Rev. Econ. Stud.*, Jan. 1969, 36, 5-12.
- J. Hicks, "Liquidity," *Econ. J.*, Dec. 1962, 72, 787-802.
- D. Levhari, and T. N. Srivasan, "Optimal Savings under Uncertainty," *Rev. Econ. Stud.*, Apr. 1969, 36, 153-64.
- B. W. Lindgren, *Statistical Theory*, New York 1965.
- B. Mandelbrot, "The Variation of Certain Speculative Prices," *J. Bus. Univ. Chicago*, Oct. 1963, 36, 394-419.
- H. Markowitz, "Portfolio Selection," *J. Finance*, Mar. 1952, 7, 77-91.
- L. J. Mirman, "Uncertainty and Optimal Consumption Decisions," *Econometrica*, Jan. 1971, 36, 177-83.
- J. W. Pratt, "Risk Aversion in The Small and in The Large," *Econometrica*, Jan. 1964, 32, 122-36.
- M. Richter, "Cardinal Utility, Portfolio Selection and Taxation," *Rev. Econ. Stud.*, Apr. 1960, 27, 152-66.
- P. A. Samuelson, "General Proof that Diversification Pays," *J. Finance Quant. Anal.*, Mar. 1967, 2, 1-13.
- , "The Fundamental Approximation Theorem of Portfolio Analysis in Terms of Means, Variances and Higher Moments," *Rev. Econ. Stud.*, Oct. 1970, 37, 537-42.
- J. E. Stiglitz, "The Effects of Income, Wealth and Capital Gains Taxation on Risk-Taking," *Quart. J. Econ.*, May 1969, 83, 263-83.
- J. Tobin, "Liquidity Preference as Behavior Towards Risk," *Rev. Econ. Stud.*, Feb. 1958, 25, 65-85.
- , "Comment on Borch and Feldstein," *Rev. Econ. Stud.*, Jan. 1969, 36, 13-14.

# Black Education, Earnings, and Inter-regional Migration: Some New Evidence

By LEONARD WEISS AND JEFFREY G. WILLIAMSON\*

The voluminous literature on the role of education in racial income differentials leaves a number of unanswered questions which are critical to the formulation of policy in northern urban ghettos. One of the most striking results using 1960 Census data is that black returns to education are erratic and much lower than for whites (see Giora Hanoch, Lester Thurow) even when some correction is made for their region of birth (see R. Weiss). These discouraging results leave migration as almost the only systematic means of improving the relative income position of blacks, and furthermore the quick gains from migration are never likely to be as great as during the 1940's and 1950's (see James Gwartney). Even the economic value of migration may be exaggerated if it is true that the most promising people migrate (see John C. Kain and Joseph Persky). In general these earlier studies, based on the 1960 Census, offer bleak prospects for elimination of black poverty, although there may still be scope for improvement through the implementation of antidiscrimination laws (see William Landes).

A possibly more hopeful view might at-

\* The University of Wisconsin, Madison. We gratefully acknowledge the research assistance of Earl Kinmonth, Richard Kaluzny, and Nancy Williamson. The paper has benefited considerably from the comments of Dennis Aigner, Glen Cain, Arthur Goldberger, Lee Hansen, Robinson Hollister, and Harold Watts. The research was supported by funds granted to the Institute for Research on Poverty at the University of Wisconsin by the Office of Economic Opportunity pursuant to the provisions of the Economic Opportunity Act of 1964. The conclusions are the sole responsibility of the authors.

tribute the weak effect of education on black income to the alleged low quality of black education provided in parts of the South, rather than entirely to discrimination in the labor market. The traditional view is that southern black migrants are "poorly educated, have high levels of unemployment and low incomes, and place disproportionate demands on welfare and public services" (Kain and Persky, p. 294). The low quality of black migrants to northern cities might be explained by a long history of systematic discrimination in the provision of public education. If interregional differences in the quality of black education account for a large part of the poverty in the northern ghetto, then the natural policy conclusions would seem to be that the North should devote more resources to southern black education (Kain and Persky). Moreover, the northern black/white income differential might be expected to diminish in the future, since a large share of the present northern black population received its education in the South while a much larger proportion of the next generation of northern blacks will have northern educations (see, for example, Richard Day). This hypothesis is not inconsistent with the research cited above, since other studies on the black education-income relationship were unable to identify the regional source of education.

The 1967 Survey of Economic Opportunity makes it possible to investigate this more optimistic hypothesis more thoroughly and to reexamine the education-

income relation for a more recent period of labor market experience. The sample consists of a Current Population Survey for 1967 augmented by a special sample drawn from low-income areas. Among other questions, the individuals surveyed were asked their place of residence at age 16. The answer should offer an excellent proxy for geographic source of education.

The present study attempts to re-evaluate these hypotheses utilizing this new data body. Section I builds upon the earlier earnings-functions studies by using more recent data and by introducing variables which identify the region of education. In Section II, we compare our results with Thurow's and report the parameter shifts which took place during the 1960's. Finally, some conclusions are presented in Section III.

### I. The Determinants of Black Male Incomes: 1967

This section estimates black earnings functions in the United States. We disaggregate into five "regional" labor markets: rural South, small-city South, medium-city South, large-city South, and non-South (primarily urban North). No breakdown of the non-South category was attempted because the rural and small-town northern black population is very small, and inter-city differences in estimated black earnings functions outside the South are not significant.<sup>1</sup> Most of the variance in environmental and school quality should be captured in these five regional classifications.

<sup>1</sup> The 1967 sample is distributed by residence as follows: 16.3 percent rural South, 2.4 percent small-town South, 10.2 percent small-SMSA South, 26.3 percent large-SMSA South, and 44.8 percent non-South. The non-South category can be disaggregated further by the following percentages: 1.0 rural and small-town, 2.0 small SMSA, and 41.8 large SMSA. The percentages who received education in the rural, small town, and small SMSA non-South would surely be even lower. Furthermore, none of the regression results reported in Table 2 are altered by including a detailed breakdown of the non-South category by location characteristics.

TABLE 1—ACHIEVEMENT OF INDIVIDUALS: STANDARDIZED YEARS OF SCHOOLING

Region and Race	Adjusted Years of Schooling for Years of Schooling Completed	
	8 years	12 years
<b>White</b>		
Urban		
Northeast	8.0	12.0
North Central	8.0	11.6
South	7.5	11.1
West	7.7	11.5
Nonurban		
Northeast	7.7	11.1
North Central	7.7	11.1
South	7.1	10.5
West	7.7	11.1
<b>Black</b>		
Urban		
Northeast	5.9	8.7
North Central	6.0	8.7
South	5.3	7.8
West	5.6	8.1
Nonurban		
Northeast	5.6	7.8
North Central	5.6	7.8
South	4.6	6.8
West	5.6	7.8

Source: Reported in Weiss, Table 1, p. 4, where it is derived from J. Coleman et al., Table 3.121.1, p. 274.

The returns to black education are estimated by controlling for the conventional variables as well as present residence and geographic source of education. Years of education by itself has long been recognized as a poor proxy for training (Gary Becker, pp. 79-88, 124-27). The Coleman Report contains a quantitative evaluation of the divergence between student "achievement" and years of formal education. Taking northeastern urban whites as the standard, the achievement variable reported in Table 1 is the number of years of schooling less the average grades behind the urban white northeasterners.<sup>2</sup> To the extent that quality of

<sup>2</sup> One of the important conclusions of the Coleman Report was that student achievement is influenced as much by the quality of educational inputs as by the socioeconomic characteristics of the home and community (see Weiss, p. 10). Regional dummy variables

schooling affects achievement, this evidence supports the common observation regarding the low quality of black schools. The data in Table 1 also suggest that regional variation in the quality of black schooling exceeds that of white. Coupling these data with the evidence of greater rates of immigration to northern cities by southern blacks than whites, it appears reasonable to appeal to regional variance of the educational and environmental quality of black training as one potential explanation for the poor association in the aggregate between black education and earnings.

The quality of training which black migrants from the South bring with them to northern cities may have a significant impact on the economic position of northern blacks. In addition to shedding light on this issue, our region-of-education variable may yield evidence about the possible decrease in interregional educational (environmental) quality differentials associated with the improvement in southern school systems and the alleged deterioration of northern ghetto schools, as well as the effect of migration on income, having controlled for the level and source of training.

The study is limited to income from work, business, or farming, for black males of working age, 20-64. We focus on male income to insure comparability with earlier studies on the economic position of the black in America, but it might also be argued that male income is the crucial variable affecting the *long-run* movement of black families out of poverty. The analysis is restricted to labor and self-employment income because such incomes are most likely to be affected by the quantity and quality of education, and by migra-

tion. Certainly, it would be inappropriate to study transfer payments or include them as a component of income, since our concern is with the ability of blacks to attain higher economic status via education and mobility, not via welfare payments. The entire black male population of working age, whether employed or not, is included in the analysis since age, origin of education, and mobility should influence income through unemployment experience as well as through wage rates.<sup>3</sup>

To accommodate the analysis in another part of this project, the variables which measure or identify age, education, region of education, and current region of residence were introduced by classifying individuals on a four-way basis, as shown in Table 2. This cross-classification produced 900 cells. The earnings functions estimated in this section of the paper introduce dummy variables which assume values of one for individuals who are members of the specific age, education, or region class. Dummy variables are also introduced in some cases to capture interaction effects; for example, dummy variables which assume values of one where an individual is a member of a specific age *and* education class. Age and education are treated as continuous variables, however, in Section II.

The earnings function initially utilized for estimation is a multiple linear regression in a single equation and should be

<sup>3</sup> The bulk of the black males age 20-64 analyzed in this paper were either employed or looking for work (92.7 percent). Only a small part of the remainder were in school (0.5 percent), in institutions (0.3 percent) or retired (0.5 percent), but a substantial group was "ill, disabled or unable to work" (5.6 percent). Some portion of those neither employed nor looking for work may have withdrawn from the labor force lacking employment opportunity, and these surely should be included in an analysis of the sources of low black incomes. It can also be argued that criminal convictions and mental and physical health problems are elements of black poverty so that these individuals rightfully belong in our analysis as well.

identifying the source of education should capture both of these effects. The evidence regarding the large regional variance in the quality and quantity of educational inputs per student is well documented.

TABLE 2—CLASSIFICATION OF INDEPENDENT VARIABLES

Age	Education	Residence at Age 16	Residence in 1967
20-24	Less than 8 years	Rural South (<2500 population)	Rural South (<2500 population)
25-29	8-11 years		
30-34	12 years		
35-39	More than 12 years	Small-town South (population 2500 to 50,000)	Small-town South (population 2500 to 50,000)
40-44			
45-49			
50-54			
55-59		Small southern <i>SMSA</i> 's (population <250,000)	Small southern <i>SMSA</i> 's (population <250,000)
60-64		Large southern <i>SMSA</i> 's (population $\geq$ 250,000)	Large southern <i>SMSA</i> 's (population $\geq$ 250,000)
		North or West	North or West

viewed as a reduced form equation incorporating both demand and supply effects (see Hanoch, p. 6). Regressions were estimated for each 1967 residence group separately and again for the entire sample. Where the entire sample was used, additional dummy variables were introduced for 1967 residences. Regressions were run with absolute income and again with *log* of income as the dependent variable. The four resulting regressions for the entire sample appear in Tables 3A and 3B. The second pair of regressions allow for a region of education-age interaction to explore the hypothesis that regional quality differences in education have changed over time. The second pair also contain an age-education interaction, because of the well-known tendency for education to affect incomes in later years more than in early years. To facilitate computation, the age classification was collapsed into four age groups in the interaction regressions.

The results of the four regressions are shown in Tables 3A and 3B. Income from work and self-employment is the dependent variable in equations (1) and (3); its natural logarithm in equations (2) and (4). The constant terms show the average income of black males with less than an eighth-grade education who were educated

in the rural South and lived there in 1967. In the first two regressions the constant refers to those of age 20-24, and in the third and fourth, to those of age 20-29.

Our conclusions from these regressions are that age, education, and residence are major determinants of black male income but that region of education has an equivocal effect. We have explored the effect of each of these variables more thoroughly using separate regressions for each region. The results appear in Tables 4 and 5.

Age has a significant and systematic effect of the expected pattern. The five regional regressions are summarized in Table 4, which shows estimated income of *nonmigrants* with 8-11 years of education (the overall median education class). Income peaks earliest in the rural South and latest in the urban North as might be expected, and for almost all age groups income increases as we move from the rural South category to the North.

As expected, our regression results show that region of residence has a very strong effect on earnings with the income advantage increasing systematically from rural South to non-South. Certainly a significant portion of this differential is due to the tendency for those with higher potential to migrate to the larger cities and the North.

TABLE 3A—REGRESSIONS RELATING INCOME FROM  
WORK AND SELF-EMPLOYMENT TO EDUCATION,  
RESIDENCE, AND AGE: BLACK MALES, 1967  
(*t*-values in parentheses)

		Equations	
		(1)	(2)
		Coefficient	
Independent Variable		Absolute Income Dependent	Log Income Dependent
Constant		287.02 (2.36)	6.045 (50.91)
Education	8-11 years	677.80 (8.34)	0.583 (7.36)
	12 years	1246.36 (12.76)	0.859 (9.01)
	>12 years	2424.38 (20.91)	0.922 (8.15)
Residence at 16	Small-town	-26.73 (-0.20)	-0.153 (-1.18)
	South		
	Southern-SMSA	140.11 (1.04)	-0.022 (-0.16)
	<250,000		
	Large-SMSA	-187.57 (-1.75)	-0.325 (-3.11)
	South		
	North	-31.26 (- .30)	-.200 (-1.97)
Residence in 1967	Small-town	415.93 (1.78)	.510 (2.23)
	South		
	Small-SMSA	804.37 (5.13)	.431 (2.82)
	South		
	Large-SMSA	1519.27 (12.57)	.662 (5.62)
	South		
	North	2232.34 (19.40)	.776 (6.91)
Age:	25-29	1150.14 (9.13)	.691 (5.62)
	30-34	1551.00 (12.51)	.799 (6.61)
	35-39	1837.86 (14.12)	.849 (6.69)
	40-44	1996.91 (15.97)	.914 (7.50)
	45-49	1910.33 (14.88)	.680 (5.44)
	50-54	1632.53 (12.30)	.454 (3.51)
	55-59	1198.35 (8.43)	-0.110 (- .79)
	60-64	572.46 (3.80)	-1.013 (-6.90)
$R^2$		.259	.104
<i>d.f.</i> = 5765-20		5745	5745

It is evident from the data in the Survey of Economic Opportunity (*SEO*) that migration rates are far greater in the rural and small-city South among the more educated groups. Our regressions control for education, but migrants undoubtedly self-select on other bases as well. This can best be shown by reference to Table 5, which compares computed incomes for persons in the overall median age and education groups by origin of education and 1967 residence. Average incomes (weighted by numbers in the *SEO* tape) for the in-migrants are shown at the bottom of each column.<sup>4</sup> The nonmigrants appear on the diagonal. They generally earn less than the immigrants to their region (shown at the bottom), even controlling for age and education. Yet, a strong interregional discrepancy still remains as can be seen by comparing the estimated incomes of the non-migrants along the diagonal.

Contrary to conventional belief and the implications of the Coleman Report, geographic source of education does *not* have the expected effect on earnings. Although none of the "residence at age 16" coefficients were significantly different from zero in the absolute income regression, in the logarithmic and the interaction regressions (equations (2), (3) and (4)) those who received their educations in the North or the large southern *SMSA*'s received significantly *lower* incomes.<sup>5</sup> Furthermore, the

<sup>4</sup> Although it attracted considerable attention during the 1960's, it should be pointed out that black migration from the North to southern cities is relatively insignificant, at least in 1967. For example, Table 5 indicates that only 39 out of 609 in-migrants to southern cities had northern origins. The estimates are for blacks aged 40-44 and with 8-11 years of schooling.

<sup>5</sup> An obvious counter-argument is that the lower incomes of persons educated in the North are due to the selective migration effect since most of the native northerners are still there and must be compared with the superior southern blacks who have migrated to the North. Yet, the negative effect of originating in the large-SMSA South holds up within most regions. The stronger negative effect in the logarithmic regression

TABLE 3B—REGRESSIONS RELATING INCOME FROM WORK AND SELF-EMPLOYMENT TO EDUCATION, RESIDENCE, AND AGE WITH INTERACTION VARIABLES INCLUDED: BLACK MALES, 1967  
(*t*-values in parentheses)

Equations			Equations		
(3)			(4)		
Coefficient			Coefficient		
Independent Variable	Absolute Income Dependent	Log Income Dependent	Independent Variable	Absolute Income Dependent	Log Income Dependent
Constant	1503.36 (7.44)	6.583 (33.19)	Residence at 16 for those age 40-49	Small-town South 120.65 (0.34)	.105 (0.30)
Education 8-11 years	381.40 (1.72)	.646 (2.96)	Small-SMSA South	482.83 (1.55)	.674 (2.20)
12 years	926.24 (4.09)	.939 (4.21)	Large-SMSA South	526.78 (2.08)	.467 (1.88)
> 12 years	701.38 (2.66)	.248 (0.96)	North	784.86 (3.18)	.507 (2.09)
Residence at 16	Small-town 72.02 (0.26)	-.106 (-.39)	Residence at 16 for those age 50-64	Small-town South -134.89 (-.39)	-.044 (-.13)
	Southern SMSA <250,000 -251.44 (-1.01)	-.488 (2.00)	Small-SMSA South	435.04 (1.44)	.548 (1.84)
	Large-SMSA South -462.28 (-2.31)	-.516 (-2.62)	Large-SMSA South	287.77 (1.15)	.130 (0.52)
	North -445.23 (-2.37)	-.461 (-2.50)	North	462.30 (1.93)	.374 (1.59)
Residence in 1967	Small-town 423.33 (1.80)	.522 (2.26)	Education for those 30-39	8-11 years 317.95 (1.14)	-.088 (-.32)
	Small-SMSA South 860.37 (5.44)	.466 (3.00)	12 years	365.89 (1.24)	-.128 (-.44)
	Large-SMSA South 1589.42 (13.05)	.687 (5.74)	> 12 years	1923.31 (5.49)	.730 (2.12)
	North 2326.31 (20.08)	.814 (7.15)	Education for those 40-49	8-11 years 262.13 (.99)	-.160 (-.61)
Age:	30-39 472.82 (1.93)	.384 (1.59)	12 years	206.52 (.70)	-.236 (-.81)
	40-49 440.12 (1.89)	.098 (0.43)	> 12 years	2595.31 (7.57)	1.074 (3.19)
	50-65 -125.13 (-0.56)	-.810 (-3.67)	Education for those 50-64	8-11 years 248.03 (.94)	.051 (0.20)
Residence at 16 for those age 30-39	Small-town -441.35 (-1.22)	-.244 (-.69)	12 years	30.60 (.10)	.049 (0.16)
	Small-SMSA South 402.07 (1.28)	.418 (1.36)	> 12 years	2752.50 (7.62)	1.426 (4.02)
	Large-SMSA South 94.28 (0.36)	.069 (0.27)			
	North 229.24 (0.92)	.031 (0.13)	$R^2$	.258	.090
			<i>d.f.</i> = 5765-36	5729	5729

TABLE 4—ESTIMATED INCOMES FOR BLACK MALES: NONMIGRANTS,  
8-11 YEARS OF EDUCATION, 1957 RESIDENCE<sup>a</sup>

Age	Rural South	Small-town South	Small- SMSA South	Large- SMSA South	North
20-24	1898	1434	1996	2131	2830
25-29	2181	1889	3071	3344	4285
30-34	2575	2556	3895	3716	4595
35-39	2525	2626	3635	4315	4927
40-44	2368	2772	3884	4433	5180
45-49	2158	2581	3415	4160	5385
50-54	2122	3051	3101	3888	5012
55-59	1773	2512	3025	3873	4207
60-64	1328	1847	2256	2761	3831

<sup>a</sup> These income estimates are based on regressions run on each of the five regions separately, following the earnings function specification of equation (1) in Table 2.

TABLE 5—ESTIMATED INCOME FOR BLACK MALES SETTING AGE AT 40-44  
AND EDUCATION AT 8 TO 11 YEARS BY RESIDENCE AT AGE 16  
(Total numbers of observations in parentheses)<sup>a</sup>

Residence at Age 16	Residence in 1967					All Observations
	Rural South	Small Town South	Small SMSA South	Large SMSA South	North	
Rural South	2368 (910)	2747 (25)	3866 (117)	4562 (357)	5253 (489)	(1898)
Small-town South	2652 (15)	2772 (103)	3943 (32)	4678 (125)	5130 (218)	(493)
Small-SMSA South	1664 (5)	3253 (2)	3883 (418)	4877 (88)	4666 (154)	(667)
Large-SMSA South	1494 (4)	2945 (5)	3891 (14)	4433 (907)	4374 (255)	(1185)
North	4236 (8)	3986 (3)	3777 (5)	4867 (39)	4489 (1467)	(1622)
In-Migrants	2436 (32)	2910 (35)	3880 (168)	4651 (609)	4947 (1116)	
All observations	(942)	(138)	(586)	(1516)	(2583)	(5756)

<sup>a</sup> See note to Table 4.

results suggest that there may have been a relative change over time in the income effect of interregional quality differences in black education. Although those aged 20–29 suffer a serious handicap from their education received in large cities, North and South, the disadvantage for those aged 30–39 is considerably less, and those aged 40–49 are better off for having been educated there.<sup>6</sup> This evidence suggests a relative deterioration over time in educational and environmental quality for blacks “educated” in large northern and southern cities, but it is presented here only as a tentative finding. Much more research remains to be completed on this issue.

Our overall conclusion must be that interregional differences in the quality of black education have relatively weak effects on earning ability, and thus southern rural blacks suffer no competitive disadvantage in urban labor markets, North or South: on the contrary, if anything it appears to be the ghetto-educated young black who suffers the competitive disadvantage. A possible explanation for this result is that other features of rural southern origin may outweigh the disadvantage of low-quality formal education there. An implication is that the geographical shift in

population can only improve black incomes by the positive impact on income from migration and by increasing the number of years of school completed by migrant’s children.

Education has a strong and consistent effect on black incomes for the sample as a whole, and for each age group in the interaction equation. The same conclusion generally holds within each of the five regions separately, but the relative gain from education is greatest in the North and least in the rural South.

## II. Full Employment and Parameter Shifts in the 1960’s

These results based on 1967 *SEO* data differ sharply from previous studies which have found little effect for black education using 1960 Census data (see Hanoch, R. Weiss), especially at high education levels. These earlier pessimistic results are perhaps best known by Thurow’s research.

In an effort to determine the source of differences between our results and those based on the 1960 Census, we applied Thurow’s model to our data. Thurow’s human-capital function has the following specification:

$$Y_{ik} = A \prod_{g=1}^n Ed_i^{b_g} \prod_{l=1}^m Ex_i^{c_l}$$

where  $g$ =education class,  $l$ =experience class,  $b$  and  $c$  are elasticities of income  $Y$  with respect to education  $Ed$  and experience  $Ex$ ,  $i$ =years of education, and  $k$ =years of experience. The experience variable is defined as the number of years in the labor force, presumably absorbing on-the-job training, from school departure or from age 18, whichever comes later. Thus, for example, a college graduate is assumed to have started work at age 22. In order to test the hypothesis that different education and experience ranges have different elasticities, Thurow divides the continuous education variable into three separate variables (less than 8, 9–12, and

---

suggests that large-city origin tends to increase inequality, perhaps by producing a larger proportion of unemployable persons, but this is pure conjecture. The large southern *SMSA*’s were initially defined as a separate group because they are so heavily concentrated in the border states. A third of the blacks in large southern *SMSA*’s in 1960 were in Washington, Baltimore, Louisville, and Wilmington, and more than half were in those cities plus the large cities of Oklahoma and Texas. We had expected that residence in these cities at age 16 would have a positive effect on earnings relative to rural southern origins.

<sup>6</sup> These results are derived by adding the interaction coefficients for each age group above 29 to the estimated “Residence at 16” coefficient. Those educated in the North or the large *SMSA* South and aged 40–49 in 1967 were in primary school between 1925 and 1941 and in secondary school between 1933 and 1945. Many of them were in World War II and entered the labor force during the war or the postwar boom.

TABLE 6—THUROW'S HUMAN-CAPITAL FUNCTION ESTIMATED ON 1967 SEO DATA<sup>a</sup>

	Weiss-Williamson Results (1967)			Thurow Results (1960)		
	All Black	North Black	South Black	All Nonwhite	North Nonwhite	South Nonwhite
Constant Term	3.7421 (0.1692)	3.8266 (0.2693)	3.8598 (0.2262)	6.6633 (0.0573)	6.9208 (0.0515)	6.5635 (0.0528)
Coefficients: Education						
$b_1$	-1.5902 (0.2221)	-1.2579 (0.3491)	-1.4728 (0.3031)	-0.6839 (0.1919)	-0.4533 (0.1725)	-0.4908 (0.1768)
$b_2$	0.0236 (0.5609)	-0.0453 (0.7998)	-0.5627 (0.8045)	-0.5664 (0.3907)	-0.7074 (0.3512)	-1.1655 (0.3599)
$b_3$	1.7718 (0.4496)	1.5086 (0.6309)	2.2222 (0.6444)	1.3348 (0.2480)	1.2235 (0.2230)	1.7374 (0.2285)
Experience						
$c_1$	0.9073 (0.2186)	1.1588 (0.3289)	0.7295 (2.9225)	-0.3897 (0.0949)	-0.3557 (0.0853)	-0.3693 (0.0875)
$c_2$	0.6942 (0.2312)	0.1777 (0.3507)	1.0891 (0.3070)	0.7881 (0.1638)	0.7365 (0.1473)	0.8712 (0.1509)
$c_3$	4.0486 (0.4137)	5.5192 (0.6935)	3.1310 (0.5177)	-0.1378 (0.1015)	-0.1089 (0.0913)	-0.2598 (0.0935)
$c_4$	-4.0490 (0.3356)	-5.1526 (0.5804)	-3.4439 (0.4106)	0.0	0.0	0.0
$\bar{R}^2$	.143	.141	.136	.89	.89	.89
SEE	2.30	2.35	2.26	.201	.181	.185

<sup>a</sup> The 1960 results are from Thurow, Table G-1, p. 188. The estimation equation is given in the text. Figures in parentheses are standard errors.

more than 12 years education). In a similar fashion, experience is divided into four variables (0-5, 6-15, 16-35, and more than 35 years experience). In summary, Thurow's human-capital function estimates income elasticities with respect to education and experience which are allowed to vary across levels of education and experience. The estimated models for 1960 and 1967 are reported in Table 6.<sup>7</sup>

<sup>7</sup> The regression model is estimated in the following form:

$$\ln Y = \ln A + b_1 \ln Ed_1 + b_2 \ln Ed_2 + b_3 \ln Ed_3 + c_1 \ln Ex_1 + c_2 \ln Ex_2 + c_3 \ln Ex_3 + c_4 \ln Ex_4,$$

where  $Ed_i$  ( $i=1, 2, 3$ ) refers to years of schooling to a maximum of 8, 12, and more than 12 years, respectively,

Thurow's 1960 results are compared with our 1967 results in Table 7, which reports computed elasticities for 1960 and 1967, for the South, North, and the United

and  $Ex_k$  ( $k=1, 2, 3, 4$ ) refers to years of experience to a maximum of 5, 15, 35, and more than 35 years respectively. As explained in the text,  $Ex_k$  is calculated by subtracting the age at which an individual started work from his current age. We assume with Thurow that an individual started work at 18 if he finished school by that age or earlier. If not, work began at school-leaving age (Thurow, pp. 72-76 and 187-88).

To calculate education and experience elasticities, the  $b$  or  $c$  coefficients are added together. For example, the elasticity for the 0-8 years educational range is  $b_1+b_2+b_3$ , and for the elasticity for the 9-12 years range is  $b_2+b_3$ , and the elasticity for the above 12 years range is  $b_3$  (Thurow, p. 187).

TABLE 7—INCOME ELASTICITIES OF EDUCATION AND EXPERIENCE FOR MALES, BY COLOR AND REGION, 1960 AND 1967

Sample	Years of Education			Years of Experience			
	0-8	9-12	12+	0-5	6-15	16-35	35+
All white (1960)	0.11	0.72	1.73	0.20	0.71	-0.09	0
All nonwhite (1960)	0.08	0.76	1.33	0.26	0.65	-0.14	0
All black (1967)	0.21	1.80	1.77	1.60	0.69	0.00	-4.05
Northern white (1960)	0.10	0.52	1.70	0.20	0.70	-0.06	0
Northern nonwhite (1960)	0.06	0.51	1.22	0.27	0.63	-0.11	0
Northern black (1967)	0.21	1.46	1.51	1.70	0.54	0.37	-5.15
Southern white (1960)	0.11	0.90	1.91	0.20	0.75	-0.21	0
Southern nonwhite (1960)	0.08	0.57	1.74	0.24	0.61	-0.26	0
Southern black (1967)	0.19	1.66	2.22	1.51	0.78	-0.31	-3.44

The 1960 figures are from Thurow Table 5-1, p. 77. The 1967 figures are calculated from Table 6.

States as a whole. The results confirm our earlier conclusions. Thurow found education elasticities lower for nonwhites for all education classes within both regions using 1960 data. The discrepancies were especially notable for the highest education classes. The 1967 sample suggests a dramatic change. Elasticities of income with respect to education have increased in *all* regions and *all* education classes. In fact, the data suggest *higher* elasticities for blacks in 1967 than for whites in 1960! The most remarkable shifts took place in the 0-8 year range where it jumps from 0.06 to 0.21. It appears that returns to black education in 1967 were at least as great as for whites in 1960.

The secular shift in the impact of the experience variable is somewhat different. Systematic changes appear to have taken place *only* for very young black males. The income elasticity with respect to experience for black males with five years' experience or less has undergone enormous improvement.

These results emphasize the potential importance of the two sources of black improvement over the last decade.<sup>8</sup> First, the full employment conditions of the 1960's are likely to explain at least some of the shift in the experience elasticity among the young; it may also explain some portion of the improved relative returns to black education. (See Hollister for some tentative evidence on this point.) But second, there may have been an independent shift in the incidence of discrimination at all education levels as well. The strong effect of secondary and even primary education on black male incomes in

<sup>8</sup> Note that the  $\bar{R}^2$  reported from the 1960 and 1967 regressions in Table 6 are quite different, Thurow's 1960 being much higher. The explanation is straightforward. We used *individual* data in estimating the 1967 regressions while Thurow uses median income data. After decomposing the 1960 sample into race- and region-specific subsamples, the unit of observation in Thurow's estimation is the median income in seventy-two male age-education classes (Thurow, p. 75). It seems unlikely that the different parameter estimates can be attributed to within-cell variation which we detect using individuals as the unit of observation.

1967 suggests that the improved opportunities in 1967 extended considerably beyond the token employment of a few black executives that was so often noted in the late 1960's and the concurrent full employment conditions. After all, besides the 1964 Civil Rights bill, twenty of the twenty-nine states which had fair employment laws on their books in 1968 had passed such laws only in the 1960's. Furthermore, we know that such legislation explains a significant portion of the state-by-state variance in wage differentials by race (see Landes).

Finally, we show in Table 8 what happens when our "educational quality" variable is appended to Thurow's human-capital function, estimated for the northern sample. Again we find evidence that blacks educated in the rural South have distinct *advantage* while those educated in the large southern or northern cities suffer significant disadvantage. Inclusion of region of education does not significantly change the education and experience coefficients.

### III. Conclusions

In this paper we have argued that the inferiority of southern black schools (especially rural schools), alleged by the Coleman Report to account for the poverty of black migrants to the North, can be discounted. Indeed, the overall effect of a northern or large southern urban ghetto environment appears to be more harmful to black economic progress than is a rural southern origin. We cannot confirm this assertion with certainty since we have been unable to eliminate wholly the selective migration effect.

Our study also seems to show a significant shift in the earnings function for blacks during much of the past decade. In particular it appears that in 1967 education generated returns to blacks that were as high as those enjoyed by whites in 1960. The great increase in the payoff from black education during the 1960's occurred at every education level and is unlikely to be explained by mere window dressing on the part of corporations and government agencies. It also appears that the elasticity

TABLE 8—THUROW'S HUMAN-CAPITAL FUNCTION WITH  $R(16)$ :  
NORTHERN BLACK MALES, 1967<sup>a,b</sup>

Variable	Coefficient	Variable	Coefficient
Constant	4.1683 (14.6108)	Education	
Residence at Age 16: $R(16)$		0-8	-1.3708 (-3.9072)
Small-Town South	-0.3695 (-1.9214)	9-12	0.0594 (0.0741)
Small Southern SMSA	-0.2037 (-0.9307)	12+	1.5304 (2.4222)
Large Southern SMSA	-0.4521 (-2.4784)	Experience	
Non-South	-0.4421 (-3.5225)	0-5	1.1466 (3.4891)
		6-15	0.2213 (0.6313)
		16-35	5.4188 (7.8140)
		35+	-5.1029 (-8.7967)

<sup>a</sup> Estimation equation identical with that reported in Table 6, except the residence at age 16 variable is added. Figures in parentheses are  $t$ -values.

<sup>b</sup>  $\bar{R}_2 = 0.145$ ;  $SEE = 2.355$ .

of income with respect to experience for young blacks increased markedly over this period of full employment.

We feel there is sufficient evidence that increased investment in black education will generate further economic progress for the black population relative to the white. In a related study we find that if the distributions of years of school completed for the five regions of education remain the same as they were for persons reaching 20 in 1962-67 but if region of education is allowed to shift as expected, more than half of the black male population will still not have completed high school in 1987. With sufficient investment channelled into improving the quality of elementary and secondary education available to ghetto-dwellers and southern blacks alike, with improved access to advanced education and continued access to employment opportunities commonly associated with such skills, we may avoid the bleak future to which a mere perpetuation of past accomplishments would seem to condemn us.

#### REFERENCES

- G. S. Becker, *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*, New York 1964.
- J. S. Coleman et al., *Equality of Educational Opportunity*, Washington 1966.
- R. H. Day, "The Economics of Technological Change and the Demise of the Sharecropper," *Amer. Econ. Rev.*, June 1967, 57, 427-49.
- J. Gwartney, "Changes in the Nonwhite/White Income Ratio—1939-67," *Amer. Econ. Rev.*, Dec. 1970, 60, 872-83.
- G. Hanoch, "Personal Earnings and Investment in Schooling," unpublished doctoral dissertation, Univ. Chicago 1965.
- R. Hollister, "Education and Income—A Study of Cross-Sections and Cohorts," in OECD, *Conference on Policies for Educational Growth*, Paris 1970, pp. 63-136.
- J. C. Kain and J. Persky, "The North's Stake in Southern Rural Poverty," in *Rural Poverty in the United States*, Washington 1967, pp. 288-310.
- W. M. Landes, "The Economics of Fair Employment Laws," *J. Polit. Econ.*, July/Aug. 1968, 76, 507-52.
- L. Thurow, *Poverty and Discrimination*, Washington 1969.
- R. D. Weiss, "The Effect of Education on the Earnings of Blacks and Whites," *Rev. Econ. Statist.*, May 1970, 52, 150-59.
- Office of Economic Opportunity, "Survey of Economic Opportunity," (SEO) conducted spring 1967, available on tape, Data Bank, Univ. Wisconsin.

# Incentive Contracts and Competitive Bidding

By DAVID P. BARON\*

An increasing segment of economic activity is taking place in nonmarket situations in which economic agents act outside the traditional markets or create markets to deal with specific resource allocation problems. One such problem involves the selection by a buyer of a contractor using a competitive bidding process. Competitive bidding is used extensively by the government for the selection of suppliers of goods and services and for the sale of resources such as offshore oil leases. Firms may use competitive bidding for the selection of certain suppliers of factor inputs and may attempt to sell certain products in markets in which competitive price quoting is the established market mechanism.

This paper is concerned with a bidding process in which a firm has an opportunity to bid on a project under the terms of an incentive contract. Incentive contracts were introduced by the federal government, primarily the Department of Defense, as an alternative to cost-plus-fixed-fee contracts. A primary objective of incentive contracts is to encourage the contractor to keep costs down, although the simplest incentive contracts may have certain adverse effects such as leading to reductions in quality in order to lower costs. The analysis herein will focus on the effect of the terms of the contract and the firm's attitude toward risk on the bid price of the firm. Bidding for a contract is inherently risky, and attitude toward risk

will be shown to have an important effect on bidding behavior. Attitude toward risk is also important in determining the response of the firm to changes in the cost of fulfilling the contract and to the terms of the contract.

The buyer of the good or service will have difficulty in assessing the most efficient supplier because the bid price is not the *ex post* cost to the buyer. For example, if two firms are bidding for a contract, the more risk-averse firm will be shown to submit the lower bid but the probability of a cost overrun will be greater if that firm is selected, *ceteris paribus*. Since the government is one of the leading proponents of incentive contracts, the buyer will be hereafter referred to as the government. In reality, the government does not select contractors solely on the basis of bid price, but in addition considers factors such as estimated delivery schedules, design features, past performance, etc. The analysis in this paper will not explicitly consider such qualitative features but such features may be important in the firm's assessment of the probability that it will be awarded the contract.

Incentive contracts have been considered by a number of authors. In a nonbidding context<sup>1</sup> K. L. Deavers and John McCall (1966) and McCall (1970) have examined the economics of incentive contracts for firms which allocate similar resources to production for either the public or the private sector. This paper relaxes

\* Associate professor of managerial economics and decision sciences, Northwestern University. Support for this research was provided by the National Science Foundation.

<sup>1</sup> McCall considers the effects of risk aversion and mentions competitive bidding in the Appendix to his paper, but the model used herein differs substantially from his.

the assumption of allocating similar resources and considers a firm which submits a bid to the government for a contract that is in addition to a fixed level of operations in the private sector. Frederick M. Scherer has studied incentive contracts in which the incentive rate is subject to negotiation apart from a bidding process. The model utilized in this paper treats the incentive rate as fixed, but the sensitivity of the bid price to the incentive rate is considered.

Frederick T. Moore has analyzed the features of negotiated contracts and reports that the shift by the Department of Defense from cost-plus-fixed-fee contracts to incentive contracts saved ten cents per dollar expended. John G. Cross has presented an empirical examination of incentive contracts which suggests that incentive contracts may not be as effective as often claimed. I. N. Fisher found that underruns were higher for fixed-price incentive contracts than for cost-plus-fixed-fee contracts, but suggested that this may be due to higher target prices than to efficiencies induced by the incentive. In addition, he found no relationship between the terms of the incentive contract and the amount of the underrun. Deavers and McCall also found no conclusive relationship between underruns and the incentive rate.

The theory of competitive bidding was apparently first treated formally by L. Friedman, and a recent bibliography has been prepared by R. M. Stark. The works most closely related to this paper are by Vernon Smith (1966), (1967), and D. L. Hanson and C. F. Menezes. Smith has considered a bidding model for the analysis of Treasury bill auctions, but his model and analysis differ from the one introduced here. Hanson and Menezes utilize a bidding model related to the one considered here and also explore certain of the same issues. The principal differences

involve the assumptions regarding the probability distribution governing the likelihood that a bid will be accepted and the extent of the economic analysis. The risk aversion results developed herein are more general than theirs, but they have introduced a new measure of risk aversion which is useful for studying the effects of the size of a contract on bidding behavior.

### I. The Model

The analysis focuses on a firm which must determine its bid price for a contract given its current level of operations in its other lines of business. The latter will be referred to as private sector operations although they may represent sales to consumers in the private or public sectors. The contract in question may be fulfilled either by the utilization of idle fixed resources and/or the acquisition of additional resources. The following notation (which parallels that of McCall) will be used:

$C$  = the cost of fulfilling the contract which may include an opportunity cost for resources utilized

$p$  = the target cost or bid price the firm submits

$\alpha$  = the target profit rate where  $\alpha p$ , ( $\alpha \geq 0$ ) denotes "target profits"

$\beta$  = the incentive profit rate where  $\beta(p - C)$  is paid to the firm if  $p \geq C$  and is paid by the firm to the government if  $p < C$ , where  $0 \leq \beta \leq 1$

$R$  = profits from private sector operations

$w$  = the firm's wealth which is assumed to affect risk aversion<sup>2</sup>

The analysis will consider the variables  $R$  and  $C$  both as deterministic and as random variables. If a variable is being referred to as random, it will be denoted by a tilde ( $\sim$ ). Initially, both  $R$  and  $C$  will be considered deterministic. The parameters  $\alpha$

<sup>2</sup> Wealth is a measure used by the firm to parameterize risk aversion, and may depend on net worth, liquid assets, or some similar measure.

and  $\beta$  are assumed to be fixed by the government.

The profit  $\pi$  from the contract if received is

$$(1) \quad \pi = \alpha p + \beta(p - C)$$

and the total wealth  $w_T$  of the firm is

$$(2) \quad w_T = w + \pi + R$$

If  $\beta=1$  and  $\alpha=0$ , the contract is a fixed-price contract, and if  $\beta=0$ , it represents a cost-plus-fixed-fee contract. Since both are special cases of the incentive contract, only the latter will be considered. If the firm submits a bid of  $p$ , let  $(1-F(p))$  denote the assessed probability that the firm will be granted the contract, where  $(1-F(p))$  is assumed to decrease in  $p$ . That is, let  $\tilde{p}_L$  be a random variable denoting the lowest price submitted by any other firm, so if  $f(\tilde{p}_L)$  is the density function,<sup>3</sup>

$$1 - F(p) = 1 - \int_0^p f(\tilde{p}_L) d\tilde{p}_L$$

is the probability that the firm's bid  $p$  is lower than the lowest bid of all other firms. If the firm is not awarded the contract, total wealth is

$$(3) \quad w_T = w + R$$

The firm is assumed to maximize the expected utility of  $EU(p)$  of total wealth<sup>4</sup>

$$(4) \quad EU(p) = [1 - F(p)]U(w + \pi + R) + F(p)U(w + R),$$

where  $U$  is a concave von Neumann-Morgenstern utility function.

The optimal bid  $\hat{p}$  satisfies the following first-order condition,

$$(5) \quad [1 - F(\hat{p})]U'(w + \pi + R)(\alpha + \beta) - f(\hat{p})[U(w + \pi + R) - U(w + R)] = 0$$

where  $U'$  denotes the first derivative. The second-order condition will be assumed to be satisfied at  $\hat{p}$ . If  $[1 - F(\hat{p})] > 0$ , then for  $f(\hat{p}) > 0$ ,  $U(w + \pi + R) > U(w + R)$  which implies that the firm bids such that the optimal profit  $\pi$  is greater than zero. Positive profit does not imply that the optimal bid price is greater than the cost  $C$  of fulfilling the contract as long as  $\alpha > 0$ , since  $\hat{p}$  may be less than  $C$  while  $\alpha\hat{p} + \beta(\hat{p} - C) > 0$ . If  $\pi \leq 0$ ,  $\hat{p}$  will be set as high as possible which is equivalent to not bidding for the contract.

If the cost of preparing a bid is denoted by  $K$ , the argument of the utility function is  $(w_T - K)$ . The firm will thus submit a bid if  $EU(\hat{p}) > U(w + R)$  where  $\hat{p}$  is optimal. Since optimal expected utility is increasing in  $w$ , an increase in  $K$  decreases  $EU(\hat{p})$  and makes it less likely that the firm will find it optimal to submit a bid.<sup>5</sup> In the following analysis the firm will be assumed to submit a bid for the contract in question. The effect of the bid preparation cost on the optimal bid price will be considered in the next section.<sup>6</sup>

<sup>3</sup> If the bid preparation cost is reimbursed by the government conditional on being awarded the contract, the cost  $K$  appears only in the term  $U(w + R - K)$ .

<sup>4</sup> Bids may have many dimensions other than price, particularly for nonstandard goods or services. Designs may differ as may projected completion dates, performance characteristics, etc. To formally consider such qualitative features, assume that the firm has already determined the qualitative features, denoted by  $q$ , of its bid and that only the probability of being awarded the contract is affected by  $q$ . Let that probability be denoted by  $F(p, q)$  where  $F(p, q) < (>) F(p, q^0)$  if the qualitative factors are judged to be favorable (unfavorable) and  $q^0$  denotes neutral qualitative factors. The first-order condition in (5) may be rewritten as  $U'(w + \pi + R)(\alpha + \beta) - (f(\hat{p}, q)/(1 - F(\hat{p}, q)))(U(w + \pi + R) - U(w + R)) = 0$  where  $f(\hat{p}, q)/(1 - F(\hat{p}, q))$  is referred to as the "hazard rate" in reliability theory. Hanson and Menezes and M. I. Kamien and N. L. Schwartz (1970), (1972) assume, in the context of a bidding model, that the hazard rate is nondecreasing in the bid price. Analogously, if the proportionate rate of increase (with respect to  $q$ ) in the probability of obtaining the contract

<sup>3</sup> The density function is assumed to satisfy  $f(\tilde{p}_L) \geq 0$ .

<sup>4</sup> The results presented herein are essentially unchanged if the firm is assumed to maximize the expected utility of profit. The expected utility of wealth was chosen as a criterion in order to highlight the effect of firm size on the bid price.

To consider the effect of attitude toward risk on the bid price, the Pratt-Arrow index of absolute risk aversion  $r_U(y) = -U''(y)/U'(y)$ , where  $U''$  denotes the second derivative, will be used to measure the degree of risk aversion. The interpretation of absolute risk aversion measured by the index  $r$  is that an increase in  $r$  for all  $y$  (from a shift in risk preferences, for example) results in an increase in the risk premium for any risk. For a risk  $\tilde{Z}$  the risk premium  $\Delta Z$  is defined by

$$(6) \quad E_Z U(w + \tilde{Z}) \equiv U(w + E_Z Z - \Delta Z)$$

where  $E_Z$  denotes expectation with respect to  $\tilde{Z}$ . The certainty equivalent of the risk  $\tilde{Z}$  is  $E_Z Z - \Delta Z$ . An increase in risk aversion as measured by the Pratt-Arrow index will next be shown to cause the firm to decrease its optimal bid price  $\hat{p}$ .

**PROPOSITION 1:** *Let  $U_1$  and  $U_2$  be two utility functions satisfying  $r_{U_1}(y) \geq (>) r_{U_2}(y)$  for all  $y$ , and let  $\hat{p}_1$  be the optimal bid for  $U_1$  and  $\hat{p}_2$  be optimal for  $U_2$ . Then  $\hat{p}_1 \geq (>) \hat{p}_2$ .*

**PROOF:**

For  $r_{U_1}(y) \geq (>) r_{U_2}(y)$  from Pratt (p. 129, equation (22))

$$(7) \quad \begin{aligned} & [U_1(x) - U_1(v)]/U_1'(x) \\ & \geq (>) [U_2(x) - U_2(v)]/U_2'(x) \quad \text{for } x > v \end{aligned}$$

Rewrite the first-order condition for  $U_2$  as

$$[1 - F(\hat{p}_2)](\alpha + \beta) - f(\hat{p}_2)[U_2(w + \pi + R) - U_2(w + R)]/U_2'(w + \pi + R) = 0,$$

and subtract the first-derivative of  $EU_1(p)$  evaluated at  $\hat{p}_2$  from this to obtain

$$\begin{aligned} & -f(\hat{p}_2) \{ [U_2(w + \pi(\hat{p}_2) + R) \\ & \quad - U_2(w + R)]/U_2'(w + \pi(\hat{p}_2) + R) \\ & \quad - [U_1(w + \pi(\hat{p}_2) + R) \\ & \quad - U_1(w + R)]/U_1'(w + \pi(\hat{p}_2) + R) \} \end{aligned}$$

is greater if the qualitative factors are unfavorable than if the qualitative factors are favorable, higher optimal bid prices result from favorable qualitative factors.

Since  $\pi(\hat{p}_2) > 0$ , let  $x = w + \pi(\hat{p}_2) + R$  and  $v = w + R$ , and the above expression is  $\leq (<) 0$  by (7). Substituting  $[U_2(x) - U_2(v)]/U_2'(x) = (1 - F(\hat{p}_2))(\alpha + \beta)/f(\hat{p}_2)$  yields (8) which implies that the first derivative of  $EU_1(p)$  at  $\hat{p}_2$  is less than or equal to (less than) zero.

$$(8) \quad \begin{aligned} & [1 - F(\hat{p}_2)](\alpha + \beta) \\ & - f(\hat{p}_2)[U_1(w + \alpha\hat{p}_2 + \beta(\hat{p}_2 - C) + R) \\ & - U_1(w + R)]/U_1'(w + \alpha\hat{p}_2 \\ & + \beta(\hat{p}_2 - C) + R) \leq (<) 0 \end{aligned}$$

Assuming that the second-order condition is negative, the bid must be decreased to drive (8) to zero, so  $\hat{p}_1 \leq (<) \hat{p}_2$ .

The result of Proposition 1 holds if the cost  $\bar{C}$  and private sector profits  $\bar{R}$  are uncertain, since taking the expectation of (5) and (8) does not change the sign of (8). The result also holds for all values of  $\alpha \geq 0$  and  $0 \leq \beta \leq 1$ , and thus applies to fixed-price and cost-plus-fixed-fee contracts.

The interpretation of Proposition 1 is that an increase in risk aversion causes the firm to lower its bid price in order to obtain a greater probability of being awarded the contract. To obtain the higher probability, the firm is willing to accept a lower profit on the contract. The lower price reduces the "risk" of not receiving the contract. The firm thus responds to an increase in risk aversion by exchanging profit for a higher probability of receiving the contract. When profit from the contract is reduced to zero, the firm will not submit a bid.

Further insight into this result may be obtained by writing expected utility in the form of (6) or

$$(9) \quad EU(p) \equiv U[w + (1 - F(p))\pi + R - \Delta\pi_{pL}],$$

where  $\Delta\pi_{pL}$  is the risk premium with respect to the random variable  $\tilde{p}_L$ . If  $U$  is linear,  $\Delta\pi_{pL} = 0$  and the first-order condition for (9) may be written as

$$(10) \quad 1 - F(\hat{p}) = f(\hat{p})(\alpha\hat{p} + \beta(\hat{p} - C))/(\alpha + \beta)$$

An increase in risk aversion makes  $\Delta\pi_{PL}$  positive for the price satisfying (10) and the firm responds by reducing its bid price which reduces profit but reduces the risk premium by a greater amount. For  $U$  strictly concave the optimal price is given by

$$(11) \quad 1 - F(\hat{p}) = \left[ f(\hat{p})(\alpha\hat{p} + \beta(\hat{p} - C)) + \frac{d\Delta\pi_{PL}}{d\hat{p}} \Big|_{\hat{p}=\hat{p}} \right] / (\alpha + \beta),$$

which implies that the marginal risk premium is positive. The marginal risk premium may be interpreted as a marginal cost resulting from risk aversion. An increase in risk aversion causes the firm to exchange some profit for a reduction in the risk premium which may be obtained by reducing the bid price.

If the government uses the bid price as a measure of the efficiency of its potential suppliers, that price will in general reflect five factors included in this model: 1) the cost  $C$ ; 2) the firm's subjective assessment of the probability of receiving the contract; 3) the firm's attitude toward risk; 4) the level of private sector profits  $R$ ; and 5) the contract parameters  $\alpha$  and  $\beta$ . Two firms which have the same cost and private sector profit and the same subjective assessment of the probabilities may submit different bid prices with the lower bid price reflecting greater risk aversion. A firm that is risk preferring will appear to be less efficient than any risk-neutral or risk-averse firm, *ceteris paribus*, if only the bid price is considered. If the government awards the contract to the firm with the lowest bid price and all firms are alike except for attitude toward risk, the likelihood of a cost overrun ( $\hat{p} < \text{ex post } \bar{C}$ ) is greater than if a firm with a higher bid is

selected. The government may thus incur a higher probability of a cost overrun. The effect of uncertainty regarding  $\bar{C}$  and  $\bar{R}$  will be considered in the next section with sensitivity to the contract terms and cost considered in Section III. Conclusions are offered in the final section.

## II. Taxes, Size Effects, and Uncertain Costs and Private Sector Profits

The firm contemplating its bid for a contract may recognize that its private sector profits are uncertain, and for certain classes of utility functions that uncertainty will affect the optimal bid price. In addition, firms, particularly those supplying technologically complex hardware or undertaking projects of long duration, may be uncertain about the cost of the good or service in question.

To consider the effect of uncertain private sector profits or costs on the optimal bid price, the effect of wealth changes must be determined. If  $r_U(y)$  is a decreasing (increasing) (constant) function of  $y$ ,  $U$  is said to exhibit decreasing (increasing) (constant) absolute risk aversion in the sense that the risk premium for any risk is a decreasing (increasing) (constant) function of  $y$ . The effect of wealth is given in the following result.

**PROPOSITION 2:** *For  $U$  exhibiting decreasing (increasing) (constant) absolute risk aversion,  $\hat{p}$  is an increasing (decreasing) (constant) function of  $w$ .<sup>7</sup>*

**PROOF:**

Let  $U_1(w) = U(w)$  and let  $U_2(w) = U(w + \Delta w)$  with  $\Delta w > 0$ . Observing that  $r_{U_1}(w) > (<) (=) r_{U_2}(w)$  for  $U$  exhibiting decreasing (increasing) (constant) risk aversion, the proof is the same as for Proposition 1.

If in (9)  $[(1 - F(p))\pi - \Delta\pi_{PL}] < 0$ , the

<sup>7</sup> Hanson and Menezes obtain a similar result by assuming that the hazard rate  $f(p)/[1 - F(p)]$  is a nondecreasing function of  $p$ .

firm will not submit a bid because utility, if the firm does not bid, will be greater than expected utility if it does bid. Since for a fixed price,  $\Delta\pi_{PL}$  is decreasing in wealth for  $U$  decreasingly risk averse, a wealthier firm is more likely to bid than a less wealthy firm. Arrow (1965) has suggested that decreasing absolute risk aversion is a reasonable assumption, so the effect of wealth on the decision to bid characterized here is an explicit demonstration of the following hypothesis of Cross: "Small firms would shy away from risk-bearing [incentive] contracts even more than large ones, being less able to bear the cost of uncertainty" (p. 211). Data presented by Cross tends to support this result. Proposition 2 further suggests that the wealthier the firm the greater the bid price with decreasing risk aversion.

By Proposition 2 an increase in the bid preparation cost  $K$  will result in a decrease (increase) (no change) in the optimal bid price for decreasing (increasing) (constant) absolute risk aversion. If the firm is decreasingly risk averse, an increase in the bid preparation cost decreases wealth which decreases the optimal bid price. This counterintuitive result is plausible when one recalls from the previous section that the submission of a bid is less likely with an increase in the bid preparation cost.

The effect on the optimal bid price of a lump sum tax (or subsidy) is determined directly from Proposition 2, since an increase in the tax (subsidy) decreases (increases) wealth. The effect of a change in the profits tax rate  $t$ , which applies to  $\pi$  and  $R$  but not  $w$ , is given by

$$(12) \quad \frac{d\hat{p}}{dt} = -(\pi + R) \frac{d\hat{p}}{dw} - (1/U_{pp})f(\hat{p})U' \cdot \left[ \pi - \left( \frac{1}{1-t} \right) \frac{U(x) - U(v)}{U'(v)} \right]$$

where  $U_{pp}$  denotes the second-order condition,  $v = w + (1-t)R$ , and  $x = w + (1-t)$

$\cdot (\pi + R)$ . From Pratt's equation 21 the term in brackets is positive, so for constant absolute risk aversion  $d\hat{p}/dw = 0$  and  $d\hat{p}/dt > 0$  indicating that an increase in the tax rate increases the optimal bid price. The last term in (12) will be referred to as the pure tax effect and the first term as the risk aversion effect. For increasing absolute risk aversion  $d\hat{p}/dw < 0$  and  $d\hat{p}/dt > 0$ , but for decreasing absolute risk aversion  $d\hat{p}/dw > 0$  and the sign of  $d\hat{p}/dt$  is unclear. For the latter case the risk effect is negative, since an increase in the income tax decreases wealth and increases risk aversion which has a decreasing effect on the optimal bid price. The risk aversion effect and the pure tax effect act in opposite directions. The effects of an income tax are summarized in the following proposition.

**PROPOSITION 3:** *An increase in the profits tax rate results in an increase in the optimal bid price if the firm has nondecreasing absolute risk aversion but may result in either an increase or a decrease for decreasing absolute risk aversion.<sup>8</sup> Ceteris paribus,  $d\hat{p}/dt$  is greater (less) for increasing (decreasing) absolute risk aversion than for constant absolute risk aversion.*

Proposition 2 may be used to indicate the effect of uncertainty regarding private sector profits. Let  $\bar{R}$  be a random variable which is independent of  $\tilde{p}_L$  and has a distribution function that is independent of the bid price and the parameters of the model. Let  $E_R R$  denote the expected value of  $\bar{R}$ . With  $C$  deterministic expected utility is

$$(13) \quad \begin{aligned} EU(p) &= [1 - F(p)]E_R U(w + \pi + \bar{R}) \\ &\quad + F(p)E_R U(w + \bar{R}) \\ &= (1 - F(p))U(w + \pi + E_R R - \Delta\pi_{R*}) \\ &\quad + F(p)U(w + E_R R - \Delta\pi_{R*}), \end{aligned}$$

<sup>8</sup> The effect of a wealth tax is given by (12) with  $(w + \pi + R)(1-t)$  replacing  $(\pi + R)$  in the first term, so a wealth tax has an effect analogous to that of a profits tax.

where  $\Delta\pi_{R^*}$  and  $\Delta\pi_{R^{**}}$  are the risk premiums which may in general be different because of the effect of  $\pi$ . For  $U$  strictly concave  $\Delta\pi_{R^*}$  and  $\Delta\pi_{R^{**}}$  are positive, so (13) is equivalent to a utility function with a lower wealth than with  $R$  constant and equal to its expected value. Uncertain private sector receipts thus is equivalent to a decrease in wealth. It follows then that the optimal bid price is greater (less) with uncertain private sector receipts than with using the mean as a point estimate for increasing (decreasing) absolute risk aversion. The firm will submit a bid only if  $\pi - \Delta\pi_{R^*} > -\Delta\pi_{R^{**}}$  at the bid price satisfying the first-order condition, since only then will utility be greater with the contract.

If the firm is decreasingly risk averse, uncertainty regarding private sector profits leads to a decreased bid price. Uncertain private sector profits increase the risk to the risk averse firm and the firm responds by reducing price in order to increase the probability of obtaining the contract. The risk represented by  $\bar{R}$  gives an additional marginal cost by affecting  $\Delta\pi_{p_L}$  which leads the firm to decrease its bid price. Uncertain private sector profits further complicate the government's task of distinguishing between efficient and inefficient firms.

Uncertainty regarding the cost of fulfilling a contract may also affect the bid price. Let  $\tilde{C}$  be a random variable independent of  $\tilde{p}_L$  with a distribution function that does not depend upon  $p$  or the parameters of the model. Letting  $E_C C$  denote the expected value of  $\tilde{C}$ , expected utility is

$$\begin{aligned} EU(p) &= E_C[[1 - F(p)]U(w + \alpha p \\ &\quad + \beta(p - \tilde{C}) + R) + F(p)U(w + R)] \\ (14) \quad &= [1 - F(p)]U(w + \alpha p + \beta(p - E_C C) \\ &\quad - \Delta\pi_C + R) + F(p)U(w + R) \end{aligned}$$

where  $\Delta\pi_C$  is the risk premium associated with uncertainty regarding  $\tilde{C}$ . The firm

will submit a bid only if  $\alpha\hat{p} + \beta(\hat{p} - E_C C) - \Delta\pi_C \geq 0$ , so uncertain costs may cause the firm to not submit a bid when it otherwise would if  $C$  was deterministic and equal to  $E_C C$ . Moore, p. 223, illustrates a similar phenomenon in the context of negotiated contracts. If the firm is risk neutral, uncertain costs do not affect the bid price.

To consider the effect of cost uncertainty, note that the second derivative of the first-order condition with respect to  $C$  for  $\bar{R}$  uncertain is

$$\begin{aligned} (15) \quad &\beta^2 E_R\{(1 - F(\hat{p}))U'''(w + \pi + \bar{R}) \\ &\quad - f(\hat{p})U''(w + \pi + \bar{R})\} \equiv g''(C) \end{aligned}$$

For nonincreasing absolute risk aversion  $U''' > 0$ , so the first-order condition is strictly convex in  $C$ . (For increasing absolute risk aversion  $U'''$  may have any sign and (15) may be positive or negative.) By Jensen's inequality  $E_{CG}(C) > g(E_C C)$ , so the first-order condition evaluated at  $E_C C$  (and  $\hat{p}$ ) becomes positive with uncertain cost with mean  $E_C C$ . Consequently, if the first-order condition is convex in  $C$ , the optimal bid price  $p^*$  under cost uncertainty is greater than  $\hat{p}$  with cost certain and equal to  $E_C C$ . Point estimation of  $\tilde{C}$  thus results in a nonoptimal bid. These results are summarized by the following proposition.

**PROPOSITION 4:** *Let  $\hat{p}$  be optimal with  $C$  deterministic and equal to  $E_C C$  and let  $p^*$  be optimal with  $\tilde{C}$  uncertain. For  $U$  exhibiting nonincreasing absolute risk aversion,  $p^* > \hat{p}$ .*

Cross has suggested a similar result.

The interpretation of Proposition 4 is that the firm requires a greater bid price when cost is uncertain in order to increase the margin between the bid and cost. This result is intuitively reasonable, since a risk-averse firm is less likely to prefer the

greater risk associated with uncertain costs. The firm thus increases its bid price and reduces the probability of obtaining the contract. The firm with uncertain costs will thus appear to be less efficient, *ceteris paribus*, if the bid price is used as a measure of *ex ante* efficiency.

### III. Sensitivity to the Contract Terms and Cost

This section investigates the response of the optimal bid price to cost changes and changes in the target profit and incentive rates. Initially, consider  $\bar{R}$  uncertain but cost  $C$  deterministic. The sensitivity of the optimal bid price to changes in  $C$  is given by implicitly differentiating the first-order condition for (13) which yields

$$(16) \quad \frac{d\hat{p}}{dC} = (1/EU_{pp}) \cdot \beta[(1 - F(\hat{p}))E_R U''(w + \pi + \bar{R}) \cdot (\alpha + \beta) - f(\hat{p})E_R U'(w + \pi + \bar{R})],$$

where  $EU_{pp}$  is the second-order condition which is assumed to be negative. The term in brackets is negative indicating that an increase in  $C$  results in an increase in the optimal bid price.<sup>9</sup> An increase in  $C$  might occur because of a change in the specifications for the contract or from exogenous factor price changes, for example.

The magnitude of the effect may be studied by separating  $d\hat{p}/dC$  into a risk aversion component and a pure cost component. Equation (16) may be rewritten as

$$(17) \quad \frac{d\hat{p}}{dC} = -\beta \frac{d\hat{p}}{dw} - \beta f(\hat{p})E_R U'(w + \bar{R})/EU_{pp}$$

The last term in (17) is always positive, and the sign of  $d\hat{p}/dw$  is given by Proposition 2. With constant absolute risk

aversion  $d\hat{p}/dw=0$  indicating that cost changes have no risk-aversion effect. For  $U$  exhibiting decreasing (increasing) absolute risk aversion,  $d\hat{p}/dw > (<)0$ , and price increases less (more) than with constant absolute risk aversion, *ceteris paribus*.

**PROPOSITION 5:** *For  $U$  exhibiting decreasing (increasing) absolute risk aversion,  $d\hat{p}/dC < (>) (d\hat{p}/dC)_*$ , where  $(d\hat{p}/dC)_*$  corresponds to  $U$  constant absolute risk aversion, *ceteris paribus*.*

The interpretation of Proposition 5 is that an increase in cost increases (decreases) the risk aversion exhibited by  $U$  for decreasing (increasing) absolute risk aversion. The risk effect thus acts in the opposite (same) direction as the cost effect for  $U$  exhibiting decreasing (increasing) absolute risk aversion.

If the contract under consideration is cost-plus-fixed-fee ( $\beta=0$ ), changes in cost  $C$  do not affect the bid price. Cost-plus-fixed-fee contracts thus do not provide any necessary relationship between the bid price and the firm's estimate of the cost of fulfilling the contract. The only factor holding down the firm's bid price is the assessed probability of the lowest bid of competitors. Cross makes a similar observation.

With  $C$  deterministic the effect of changes in the target profit rate  $\alpha$  and the incentive rate  $\beta$  may be related to the expression in (16). It is easily shown that

$$(18) \quad \frac{d\hat{p}}{d\beta} = -((\hat{p}-C)/\beta) \frac{d\hat{p}}{dC} - (1-F(\hat{p}))E_R U'(w + \pi + \bar{R})/EU_{pp}$$

$$(19) \quad \frac{d\hat{p}}{d\alpha} = -(\hat{p}/\beta) \frac{d\hat{p}}{dC} - (1-F(\hat{p}))E_R U'(w + \pi + \bar{R})/EU_{pp}$$

The last term in (18) and (19) is positive

<sup>9</sup> The same results obtain with cost uncertain if an increase in cost is taken to be equivalent to a positive translation in the distribution of  $\bar{C}$ .

and may be interpreted as a conditional (on winning the contract) marginal utility due to an increase in  $\alpha$  or  $\beta$ . In (19) the first term is negative while the first term in (18) is positive if  $\hat{p} - C < 0$  and negative if  $\hat{p} - C > 0$ . An increase in the target profit rate may lead to either an increase or a decrease in the bid price. The firm will exchange a portion of the additional profit due to an increase in  $\alpha$  for a greater probability of winning the contract unless the conditional marginal utility is greater than the first term in (19). If the bid price is less than or equal to the cost  $C$ , an increase in the incentive rate results in a higher bid price, since the firm desires to decrease the negative margin ( $\hat{p} - C$ ) at the expense of a lower probability of receiving the contract. If the bid price is sufficiently greater than cost, the bid price may be decreased by an increase in the incentive rate. The firm then is willing to exchange profit for a greater probability of being awarded the contract.

The response of the optimal bid price to changes in the incentive rate is particularly important when the firm is uncertain about cost. Scherer indicates that in negotiated contracts risk aversion causes contractors to prefer low values of  $\beta$ . With  $\beta$  fixed by the government in a competitive bidding situation, one might expect that the risk-averse firm would increase its bid price in order to provide a greater margin between bid and cost. Fisher and Cross present empirical data to support the hypothesis that as  $\beta$  increases fewer cost overruns occur and suggest that higher target costs (bid prices) might be the cause. Deavers and McCall found no conclusive relationship between  $\beta$  and the proportion of overruns. G. J. Feeney, W. H. McGlothlin, and R. J. Wolfson conducted a laboratory experiment involving bidding for incentive contracts which indicated that "... both the average bids and expected profit increased monotonically with the sharing rate."

To study the impact of  $\beta$  on the optimal bid price when cost is uncertain, consider  $R$  deterministic and denote the risk premium associated with cost as  $\Delta\pi_c(\alpha, \beta, \hat{p})$  where the dependence on  $w + R$  is not explicitly denoted. The risk premium is defined by

$$\begin{aligned} E_c U(w + \hat{\pi} + R) \\ (20) \quad & \equiv U(w + \alpha\hat{p} + \beta(\hat{p} - E_c C) \\ & - \Delta\pi_c(\alpha, \beta, \hat{p}) + R), \end{aligned}$$

and  $\alpha\hat{p} + \beta(\hat{p} - E_c C) - \Delta\pi_c(\alpha, \beta, \hat{p}) \equiv CE\pi$  is the certainty equivalent profit if the contract is won. Using (20) in (4) and differentiating the first-order condition implicitly yields

$$\begin{aligned} \frac{d\hat{p}}{d\beta} = & -(1/EU_{pp}) \\ (21) \quad & \cdot [\{ (1 - F(\hat{p})) U''(w + CE\pi + R) \\ & \cdot (\alpha + \beta - d\Delta\pi_c(\alpha, \beta, \hat{p})/d\hat{p}) \\ & - f(\hat{p}) U'(w + CE\pi + R) \} (\partial CE\pi/\partial\beta) \\ & + (1 - F(\hat{p})) U'(w + CE\pi + R) \\ & \cdot \partial^2 CE\pi/\partial\beta\partial\hat{p}] \end{aligned}$$

where  $EU_{pp}$  is the second-order condition for  $\hat{C}$  uncertain. It is easily shown that

$$\begin{aligned} U''(w + CE\pi + R) \\ \cdot (\alpha + \beta - d\Delta\pi_c(\alpha, \beta, \hat{p})/d\hat{p}) < 0 \end{aligned}$$

and thus the term in  $\{ \}$  in (21) is negative. For  $U$  constant absolute risk averse  $\partial^2 CE\pi/\partial\beta\partial\hat{p} = 1$  and  $d\hat{p}/d\beta$  is positive if  $\partial CE\pi/\partial\beta = \hat{p} - E_c C - \partial\Delta\pi_c(\alpha, \beta, \hat{p})/\partial\beta < 0$ . That is, if an increase in the incentive rate reduces the certainty equivalent, the optimal bid price increases. Since an increase in  $\beta$  means that the firm has a greater share of cost over- or underruns, it seems reasonable that  $\partial CE\pi/\partial\beta < 0$  for a risk-averse firm. The constant absolute risk-averse firm thus seeks a higher profit margin by increasing the optimal bid price in exchange for a lower probability of winning the contract. This result supports the ob-

servations of Scherer, Cross, Fisher, and Feeney, McGlothlin, and Wolfson.

As an example, consider the constant absolute risk-averse utility function  $U(y) = -\exp(-\gamma y)$ ,  $\alpha > 0$ , and assume that  $\tilde{C}$  is normally distributed with mean  $E_C C$  and variance  $\sigma^2$ . It may be shown that

$$(22) \quad \begin{aligned} E_C U(w + \alpha p + \beta(p - \tilde{C}) + R) \\ = U(w + \alpha p + \beta(p - E_C C) \\ + R - (\gamma/2)\sigma^2\beta^2) \end{aligned}$$

where the certainty equivalent profit is  $CE\pi = \alpha p + \beta(p - E_C C) - (\gamma/2)\sigma^2\beta^2$ . The risk premium is increasing in  $\beta$  indicating that the firm absorbs a greater share of the risk associated with  $\tilde{C}$ . Equation (21) for this example is

$$(23) \quad \begin{aligned} \frac{d\hat{p}}{d\beta} = & - (1/EU_{pp}) [\{ (1 - F(\hat{p})) \\ & \cdot U''(w + CE\pi + R)(\alpha + \beta) \\ & - f(\hat{p})U'(w + CE\pi + R) \} \\ & \cdot (\hat{p} - E_C C - \gamma\sigma^2\beta) \\ & + (1 - F(\hat{p}))U'(w + CE\pi + R)], \end{aligned}$$

where  $\hat{p} - E_C C - \gamma\sigma^2\beta = \partial CE\pi / \partial \beta$ . If  $\hat{p} - E_C C - \gamma\sigma^2\beta < 0$ , the optimal bid price is increasing in the incentive rate. For this example, the optimal bid price is also an increasing function of the variance of cost.<sup>10</sup>

The government's stated goal (see Moore, p. 220) is for the target price to accurately reflect expected cost or  $p = E_C C$ . Note that if this holds at the optimal bid price then  $\partial CE\pi / \partial \beta < 0$  if  $\partial \Delta\pi_c(\alpha, \beta, \hat{p}) / \partial \beta > 0$ . For the above example,  $\partial \Delta\pi_c(\alpha, \beta, \hat{p}) / \partial \beta = \gamma\sigma^2\beta > 0$ , so an increase in  $\beta$  always results in a higher bid price. Fisher suggests that  $\hat{p} > E_C C$  but  $\partial CE\pi / \partial \beta$  still is likely to be negative. The *ex post* cost to the government is  $\pi + C = \alpha\hat{p}_m + \beta(\hat{p}_m - C) + C$  where  $C$  is the

actual cost of the project and  $\hat{p}_m$  is the minimum of all bids.<sup>11</sup> If an increase in the incentive rate increases the optimal bid price for all firms,  $\hat{p}_m$  increases as does  $\pi$ . For incentive contracts to be desirable from the government's point of view, the cost reduction induced by the incentive will have to be larger than the increase in  $\pi$  caused by increased bid prices.<sup>12</sup>

#### IV. Conclusions

The effect of risk aversion on the optimal bid price is essentially to decrease the bid price and thus the most risk-averse firm, *ceteris paribus*, will appear to be the most efficient if the bid price is used as a measure of efficiency. If firms have the same costs, utility function, and initial wealth level, the firm with the most risky (as measured by the risk premium) private sector profits will have the lowest bid price with decreasing absolute risk aversion. With decreasing absolute risk aversion, an increase in the profits tax rate may increase or decrease the optimal bid price, while an increase in bid preparation costs decreases the optimal bid price. If the firm has nonincreasing absolute risk aversion, uncertain costs of fulfilling the contract cause the firm to increase its bid price in order to provide a cushion for possible cost overruns. The increase in the bid price is greater with increasing absolute risk aversion than with decreasing absolute risk aversion. Increases in the incentive rate are likely to result in higher bids which may mean that incentive contracts may not be better for the buyer than cost-plus-fixed-fee contracts.

The analysis herein may be generalized in a number of ways. An individual firm may have a number of contracts for which

<sup>11</sup> Arrow (1966) has suggested that the government should be risk neutral in its procuring activities and that suggestion is used here, although Cross suggests that government officials may actually act in a risk-averse manner.

<sup>12</sup> Changes in the incentive rate may have other effects such as bringing more (or fewer) firms into competition for the contract.

<sup>10</sup> The effect of "increasing uncertainty" regarding cost may be studied more generally using the analysis suggested by M. Rothschild and J. E. Stiglitz.

it may submit bids, and its decision regarding bid prices is essentially a portfolio selection problem. The game theoretic aspects involving the reactions of the other firms which might also bid on a particular contract may also be important. The analysis here is thus essentially static and myopic in that the firm is assumed to ignore the effects of its decision making on future opportunities to bid. J. H. Griesmer and M. Shubik (1963a, b, c), R. B. Wilson (1967) (1969), and I. H. LaValle have approached bidding processes from a game theoretic point of view. The model also may be generalized by considering constraints on the profit of the firm and adjustments or renegotiation possibilities once the contract has been let. In addition, the effect of incentives attached to qualitative features and activities within the project is a potential area for study.

## REFERENCES

- K. J. Arrow, *Aspects of the Theory of Risk Bearing*, Helsinki 1965.
- , "Discounting and Public Investment Criteria," in A. V. Kneese and S. C. Smith, eds., *Water Research*, Baltimore 1966.
- J. G. Cross, "A Reappraisal of Cost Incentives in Defense Contracts," *Western Econ. J.*, June 1968, 6, 205-25.
- K. L. Deavers and J. J. McCall, *Notes on Incentive Contracting*, The RAND Corp., RM-5019-PR, Sept. 1966.
- G. J. Feeney, W. H. McGlothlin, and R. J. Wolfson, *Risk Aversion in Incentive Contracting: An Experiment*, The RAND Corp., RM-4231-PR, Aug. 1964.
- I. N. Fisher, *A Reappraisal of Incentive Contracting Experience*, The RAND Corp., RM-5700-PR, July 1968.
- L. Friedman, "A Competitive-Bidding Strategy," *Operations Res.*, 1956, 4, 104-12.
- J. H. Griesmer and M. Shubik, (1963a) "Toward a Study of Bidding Processes: Some Constant-Sum Games," *Naval Res. Logs. Quart.*, Mar. 1963, 10, 11-21.
- and ———, (1963b) "Toward a Study of Bidding Processes, Part II: Games with Capacity Limitations," *Naval Res. Logs. Quart.*, June 1963, 10, 151-73.
- and ———, (1963c) "Toward a Study of Bidding Processes, Part III: Some Special Models," *Naval Res. Logs. Quart.*, Sept. 1963, 10, 199-217.
- D. L. Hanson and C. F. Menezes, "Risk Aversion and Bidding Theory," in J. P. Quirk and A. M. Zarley, eds., *Papers in Quantitative Economics*, Lawrence 1968, 521-42.
- M. I. Kamien and N. L. Schwartz, "Revelation of Preference for a Public Good with Imperfect Exclusion," *Publ. Choice*, fall 1970, 9, 19-30.
- and ———, "Exclusion Costs and the Provision of Public Goods," *Publ. Choice*, forthcoming spring 1972.
- I. H. LaValle, "A Bayesian Approach to an Individual Player's Choice of Bid in Competitive Sealed Auctions," *Manage. Sci.*, Mar. 1967, 13, 584-97.
- J. J. McCall, "The Simple Economics of Incentive Contracting," *Amer. Econ. Rev.*, Dec. 1970, 60, 837-46.
- F. T. Moore, "Incentive Contracts," in S. Enke, ed., *Defense Economics*, Englewood Cliffs 1967, ch. 12.
- J. W. Pratt, "Risk Aversion in the Small and in the Large," *Econometrica*, Jan./Apr. 1964, 32, 122-36.
- M. Rothschild and J. E. Stiglitz, "Increasing Risk II: Its Economic Consequences," *J. Econ. Theor.*, Mar. 1971, 3, 66-84.
- F. M. Scherer, "The Theory of Contractual Incentives for Cost Reduction," *Quart. J. Econ.*, May 1964, 78, 257-80.
- V. L. Smith, "Bidding Theory and the Treasury Bill Auction: Does Price Discrimination Increase Bill Prices?" *Rev. Econ. Statist.*, May 1966, 48, 141-46.
- , "Experimental Studies of Discrimination versus Competition in Sealed-Bid Auction Markets," *J. Bus., Univ. Chicago*, Jan. 1967, 40, 56-84.
- R. M. Stark, "Competitive Bidding: A Comprehensive Bibliography," *Operations Res.*, Mar./Apr. 1971, 19, 484-90.
- R. B. Wilson, "Competitive Bidding with Asymmetric Information," *Manage. Sci.*, July 1967, 13, 816-20.
- , "Competitive Bidding with Disparate Options," *Manage. Sci.*, Mar. 1969, 15, 446-48.

# COMMUNICATIONS

## Price Discrimination by Regulated Motor Carriers

By JOSEPHINE E. OLSON\*

"Value-of-service pricing" has been traditional in many forms of freight transportation. The American railroads applied value-of-service pricing almost as soon as they began to operate; the practice was continued under the Interstate Commerce Act of 1887 and long after the railroads had lost their monopoly on inland transportation. When regulation under the Interstate Commerce Act was extended to a large part of the motor carrier industry in 1936, value-of-service pricing became a declared policy of motor carrier rate making as well. In fact the motor carriers adopted the rail class rate structure almost intact in their first rate publications under regulation.

To the economist, value-of-service pricing means price discrimination. A firm with monopoly power and the ability to separate demand for its product into separate markets can increase its profits by charging higher rates to its customers with less elastic demands. Profits are maximized when the marginal cost of total output is equated with marginal revenue in each separate market. It is the purpose of this paper to extend the simple one-product model of price discrimination to the motor common carrier freight industry in order to predict its rate structure under the assumption of monopoly price discrimination. The model is then tested using class rates of motor common carriers of general freight.

\* Assistant professor of business administration, Graduate School of Business, University of Pittsburgh. Work for this study was begun under a National Science Foundation fellowship. I would like to thank G. H. Borts, Benjamin Chinitz, Mark B. Schupack, and Albert Zucker for their assistance.

### I. The Motor Carrier Industry

Before developing the model it is useful to describe the motor common carrier industry and the conditions which allow price discrimination.

Without the existence of regulation the motor carrier industry would appear to be one of the best examples of a perfectly competitive industry. Its product is the movement of goods between two points within a given time period. Some differences in quality of service may be possible (for example, differences in speed of delivery and in reliability), but they are relatively minor and difficult to maintain.<sup>1</sup> There are few economies of scale in the industry. The initial amount of capital needed to enter the industry is small and the output-capital ratio is high. Much of the invested capital is in equipment such as trucks, tractors, and trailers, which are depreciated over a short number of years. Rapid adjustment of output to shifts in the composition of demand is facilitated by the existence of markets for secondhand and rented equipment.

A few complications do exist, however, to distinguish even the unregulated motor carrier industry from the simple picture of a perfectly competitive industry. The motor carrier industry is really a multi-product industry and the simple competitive solution of price equal to marginal cost will not hold where joint products exist. There is one major joint product in the motor carrier industry, the backhaul. A shipment is made in one direction but the motor carrier vehicle must make a round trip. Although pricing is

<sup>1</sup> See Richard Farmer, p. 400.

more complicated with the existence of joint products, it can nevertheless be resolved efficiently under perfect competition.<sup>2</sup>

Although the motor carrier industry would appear to be the paragon of perfect competition, regulation by the Interstate Commerce Commission prevents two sections of the industry from operating in a perfectly competitive manner. These are contract carriers and common carriers. Contract carriers, which as their name implies are under contract to certain shippers, must have permits from the ICC and must publish their rates. Subject to even more regulation are common carriers—those carriers which offer to haul articles (within large categories such as household goods or general freight) for the general public, usually on regular schedules between particular points. Common carriers have the routes over which they operate, the commodities they carry, and the rates they charge controlled by the ICC. They cannot even exist without an operating certificate from the Commission and since new ones are difficult to obtain, entry is limited. Not under general Commission regulation are the private haulers (that is, carriers owned by firms whose primary business is something other than hauling) and the so-called “exempt haulers,” which carry agricultural products.

The share of total traffic carried by the two regulated groups of motor carriers is not insignificant. The ICC estimates that in 1965 the total amount of intercity freight traffic was 1,895.2 billion ton-miles; of these 388.4 billion or 20.5 percent were carried by motor carriers. Federally regulated motor carriers carried 140.3 billion ton-miles or 36.1 percent of all intercity motor freight traffic.<sup>3</sup> The Commission does not break down traffic between common and contract carriers for 1965; however, for 1959 it was estimated that 31.4 percent of intercity motor carrier freight traffic (the major part of regulated motor traffic) was moved by common carriers.<sup>4</sup>

Common carriers may publish their own rate lists (called tariffs) subject to ICC approval but most do so through a tariff publishing agent. The most common agent is a motor carrier conference composed of a number of member carriers, usually organized on a state or regional basis. The conference rates are used by almost all carriers operating in the applicable area, even by nonmembers. It is the existence of these rate conferences along with the regulatory powers of the Commission which allow rates to be set at other than perfectly competitive levels. Although a carrier may theoretically publish its own rates, the conference can call upon some rather powerful weapons of the Commission (especially suspension of rates and minimum rate orders) to force a recalcitrant carrier back into the price discriminating fold. Since many studies have been made documenting the combined powers of the rate conference and the ICC to prevent independent action on the part of a carrier, the matter need not be discussed further here.<sup>5</sup>

A brief explanation of the rate system reveals the almost unlimited opportunities to charge discriminatory rates. There are three major categories of common carrier rates—minimum charges, class rates, and commodity rates. The minimum charge applies on small shipments and is the minimum rate for which a shipment will be moved regardless of its weight. A class rate is a rate that can be determined from a general table of rates once the rate-making distance, weight and “class” rating of a shipment is determined. Class ratings are found in a book called a classification tariff which groups thousands of commodities into a limited number of classes to facilitate the application of rates. Even in a competitive system a number of classes might be justified because costs of hauling vary with density and other characteristics of the commodities, but in the classification tariff used by most carrier conferences,<sup>6</sup> cost of hauling and value of service

<sup>2</sup> See Tjalling Koopmans; and Walter Miklius and Daniel DeLoach, p. 937.

<sup>3</sup> ICC, *Annual Report 1967*, p. 57.

<sup>4</sup> See Walter Oi and Arthur Hurter, p. 11.

<sup>5</sup> See, for example, Walter Adams, R. C. Fellmeth, James Nelson (1965), and Robert Nelson.

<sup>6</sup> See the *National Motor Freight Classification*.

(that catchall for discrimination) are equally important criteria for classification. Even more refined discrimination is possible by the use of commodity rates on specific commodities between specific points.

The ability of motor common carriers to act as one unit in setting rates and their ability to separate their markets are alone not enough to permit monopoly price discrimination. A further condition is that there be no good substitutes for motor common carriers. For many shippers this is the case. Rail and water carriers cannot reach many of the points serviced by motor carriers. For some shippers no licensed contract carriers are available and the cost may be too high for private trucking to be a viable alternative. Potential competitors such as railroads may be excluded from effective competition if they are forced to charge the same rate as motor carriers and if their service is inferior; given a choice of rail or motor transport at the same rate, shippers would usually prefer the faster truck service.

Price discrimination may also explain the structure of class rates even if there is considerable competition. This is possible if class rates are formulated under the assumption of no competition and then where this assumption is violated, competition is met by the creation of exception ratings (where a product receives a lower class rating under certain specific conditions) or commodity rates. Some evidence for this point can be found in Commission decisions. The Commission has argued that competition between transport agencies or between shippers in different locations should not enter into consideration in the classification of goods. In *Freight Forwarder Traffic in Official Territory*, the Commission stated:

... competition is not normally a consideration in the initial establishment of classification ratings, nor is competition a normal consideration in making amendments to the classification. The basic ratings in the classification reflect maximum reasonable levels, as distinguished from exceptions ratings or lower levels which all carriers are privileged to

maintain provided they are not unduly low. [p. 69]

## II. The Price Discrimination Model

Demand for transportation is a derived demand which depends on the demand and cost conditions for the shippers' products. To derive the demand for transportation it is assumed that shippers are perfectly competitive.<sup>7</sup>

Demand for the shippers' product in each mileage ring (market) from the shippers' plants is represented as:

$$(1) \quad P_{ij} = (x_{ij}T_j)^{-1/\alpha_j}$$

where  $P_{ij}$  is the price of the product  $j$  per hundredweight in each mileage ring  $i$ ;  $x_{ij}$  is the number of shipments of product  $j$  to each ring per unit of time; and  $T_j$ , the shipment size in hundredweights (cwts.), a value which is determined exogenously.<sup>8</sup> The coefficient  $\alpha_j$  represents the absolute value of the elasticity of demand; it is a constant greater than unity.

In order to simplify the analysis it is assumed that the marginal cost of producing the product is a constant,  $c_j$ , per cwt. regardless of the firm or the level of output. The only other cost to the shippers is the transport rate to each ring,  $r_{ij}$ , per cwt.

Given the assumption of perfect competition and constant costs of production, the shippers maximize their profits when marginal cost equals the price in each market

$$(2) \quad P_{ij} = c_j + r_{ij}$$

or substituting for  $P_{ij}$ ,

$$(3) \quad (x_{ij}T_j)^{-1/\alpha_j} = c_j + r_{ij}$$

From this profit-maximizing relation the

<sup>7</sup> The basic price discrimination model is not changed by assuming shippers act as monopolies. However, interpretation of the empirical results is significantly different as will be discussed later.

<sup>8</sup> In my earlier study, pp. 44-46, I attempted to solve for shipment size considering costs of storage, the interest rate, and ordering costs as well as motor rates. A direct relation between shipment size and these variables could not be found.

demand for transportation to each market for each product is derived:

$$(4) \quad x_{ij} = \frac{1}{T_j} (r_{ij} + c_j)^{-\alpha_j}$$

The transport rates which maximize motor carrier profits depend not only on the demands for transportation but on the costs of hauling the products.

There are several problems in determining the marginal cost of hauling a shipment. A one-way shipment is essentially a joint product along with a backhaul shipment. Only if one assumes demand for transportation and the hauling characteristics of commodities are the same in both directions is one justified in treating the marginal cost of a shipment as half the round trip marginal cost. Such an assumption is made. Since the motor carrier industry is a constant cost industry, it is correct to assume the marginal cost of hauling a given commodity of a given shipment size between two points is independent of the total number of other shipments and constant in the long run.

The cost of hauling a commodity  $j$  of a given total weight  $T_j$  to mileage ring  $i$ , per cwt., is represented by the symbol  $K_{ij}$ .

Given its demand and cost conditions the motor carrier industry can maximize its total profits by acting as a discriminating monopolist setting a separate rate in each mileage ring and for each product with a different elasticity of demand or a different cost of hauling.

Total profits,  $\Pi$ , of the motor carrier industry are represented as:

$$(5) \quad \Pi = \sum_i \sum_j (r_{ij} x_{ij} T_j - K_{ij} x_{ij} T_j)$$

Substituting the shippers' demand functions for transportation (equation (4)) yields:

$$(6) \quad \Pi = \sum_i \sum_j (r_{ij} + c_j)^{-\alpha_j} (r_{ij} - K_{ij})$$

Because the marginal cost of hauling a shipment of a given product is independent of

other shipments and constant in the long run, profits are maximized by separately maximizing profits for each product and for each mileage ring:

$$(7) \quad \frac{\partial \Pi}{\partial r_{ij}} = -\alpha_j (r_{ij} + c_j)^{-\alpha_j-1} (r_{ij} - K_{ij}) + (r_{ij} + c_j)^{-\alpha_j} = 0$$

Solving for  $r_{ij}$  yields:

$$(8) \quad r_{ij} = \frac{\alpha_j}{\alpha_j - 1} K_{ij} + \frac{c_j}{\alpha_j - 1}$$

The profit maximizing transport rate for a given commodity, weight and distance is a function of the cost of hauling the commodity, its marginal cost of production, and its price elasticity of demand. Other things being equal, the transport rate increases with an increase in transport costs, with an increase in the cost of producing the good shipped, or with a decrease in the elasticity of demand for the good.

### III. Implications of the Hypothesis

The implications of this hypothesis for the motor carrier class rate structure are perhaps better seen if one refers to the ratio of the transport rate to cost of hauling:

$$(9) \quad \frac{r_{ij}}{K_{ij}} = \frac{\alpha_j}{\alpha_j - 1} + \frac{1}{\alpha_j - 1} \frac{c_j}{K_{ij}}$$

From this it can be seen that the ratio of rate to cost is not a constant but falls as the cost of shipping a given product increases due to an increase in distance or a decrease in the total size of shipment. Thus this model of monopoly price discrimination predicts that for a given product the percentage markup of the motor carrier rate over the actual cost of hauling per cwt. will be higher for short hauls than for long hauls, and higher for truckload (large) shipments than for less-than-truckload (small) shipments.

The model's prediction of a markup that decreases with length of haul is an interesting one because it offers a different explanation for this phenomenon than is usually

found. It has been argued that motor carriers have met rail competition by simply adopting rail carrier rates. Given the motor carriers' superior service characteristics, they have been able to divert much of the original rail traffic but they have ignored their own cost characteristics in doing so.<sup>9</sup> Motor carrier costs are less than rail costs on short hauls but increase more rapidly than rail costs as distance increases. If rail rates bear any resemblance to rail costs, when compared to motor carrier costs they will show a higher markup for short hauls than for long hauls.

Another variation of this explanation is the idea of umbrella rate making. This theory explains that the ICC encourages motor carriers and rails to set their rates at a level such that they can share the market rather than at levels which reflect their "inherent" cost advantages. In order to share the market the rates must be at the level of the high cost carrier.<sup>10</sup> Since the high cost carrier is the railroad for short hauls and the motor carrier for long hauls, "umbrella" rates also mean a higher percentage markup for motor rates on short hauls.

While it is not possible to reject the above two hypotheses, the price discrimination hypothesis does have several theoretical advantages over them. It explains why there may be a higher markup for short hauls even for products, shipment sizes, and areas where there is no rail competition, and it credits the motor carriers with more rational pricing behavior than blind adoption of rail class rates. It also indicates that what appears to have been motor carrier adoption of rail-oriented rates may have been rail acceptance of motor-oriented rates. Since class rates are very important for motor carriers and quite unimportant for railroads, it seems logical that motor carriers would take more interest in class rate structure than railroads.

#### IV. Testing the Hypothesis

One method of testing the price discrim-

ination hypothesis is to take a sample of commodities and to obtain information on their class rates,<sup>11</sup> on their costs of production, and on the costs of hauling them for different distances and shipment sizes. If the elasticities of product demand are randomly distributed, a linear regression of the class rates on costs of producing the commodities and on the costs of hauling them should yield an equation where the regression coefficient for cost of hauling is an estimate of  $\alpha/(\alpha-1)$  for the average value of  $\alpha$  and where the cost of production coefficient is an estimate of  $1/(\alpha-1)$ . Since  $1/(\alpha-1)$  is equal to  $\alpha/(\alpha-1)-1$ , it is logical to constrain the regression coefficients so that the coefficient of the cost of producing the commodity is one less than the coefficient of the cost of hauling. As the price elasticity of demand,  $\alpha$ , is assumed to be greater than unity, the regression coefficient of the cost of hauling should be greater than unity and the regression coefficient of the cost of producing the commodity should be positive but smaller by unity and between zero and one if  $\alpha$  is greater than two. The model does not predict an extremely high multiple correlation coefficient as the  $\alpha_j$ 's are assumed to vary.

The regression equations have been run using two separate samples of 66 commodities taken from the 1967 *National Motor Freight Classification*. The samples are not random because of the difficulty in finding information on product costs per cwt. Actually, price of the commodity per cwt. has to be used and this introduces two possible biases. One is that price of the commodity includes the transportation charges. The effect is to reduce the value of both the regression coefficients, thus underestimating the true relation between rates, cost of hauling and cost of the product, and overestimating

<sup>11</sup> Because literally millions of specific commodity rates exist, only general class rates are examined. Class rates accounted for 45.9 percent of shipments and 47.5 percent of the weight in the Middle Atlantic Region in 1963. Minimum charge shipments accounted for 43.5 percent of the shipments and 3.8 percent of the weight, and commodity rate shipments for 8.5 percent of the shipments and 46.6 percent of the weight. (See *Class Rates*, p. 2.)

<sup>9</sup> See Nelson (1960), p. 499.

<sup>10</sup> See E. W. Williams, Jr., p. 178.

the average elasticity of demand.<sup>12</sup> A second bias occurs if shippers do not sell in competitive markets. If the producer of a good also has monopoly powers, the price of his good can be raised considerably above the marginal cost of production and transportation. Here again using price rather than marginal cost reduces the size of the regression coefficients and underestimates the true relation among the variables.<sup>13</sup>

Most of the commodity prices are taken from export statistics of the Census Bureau, but a few are from the Wholesale Price Statistics of the Department of Labor.

Data on the cost of hauling average shipments of different weights and distances have been developed using 1966 ICC cost statistics for the Middle Atlantic Region, assuming costs are 100 percent variable in the long run. Since density is the most important factor affecting the cost of hauling, these average costs have been adjusted by the Commission's density adjustment ratios and are assigned to commodities according to their rating in the New England Motor Rate Bureau's classification tariff, *Coordinated Freight Classification*, No. 14. The New England ratings are used as a proxy for the relative costs of hauling different commodities because information on density and other cost-of-hauling characteristics is not available for most commodities. The New England class rating system is different from that used by the rest of the country's motor carriers. The primary basis for New England class rating is the density of a commodity, although the rating may be adjusted for other cost considerations. For example, a good may have the same density as another and yet have a higher rating if it is very expensive and therefore more liable to be stolen than the other commodity. It is assumed that the New England ratings ac-

count for all the cost-increasing aspects of high-valued commodities. If, in addition, there are elements of price discrimination in the New England ratings, then use of the New England ratings as a proxy for relative cost of hauling means high multicollinearity between the two independent variables and poor independent estimates of the coefficients.

The rates used in this study are Middle Atlantic Conference general class rates in effect in 1966 for less-than-truckload shipments between 2,000 and 4,999 pounds billed weight and for truckload shipments under 30,000 pounds billed weight for distances of 100 and 260 rate-making miles.

Three regression equations were run on both groups of 66 commodities and the results are shown in Table 1. The first regression equation for each group of commodities represents a linear regression with no constraints on the coefficients. In the second regression equation the coefficients are constrained so that the commodity price coefficient is one less than the cost-of-hauling coefficient. The third regression is, in addition, constrained through the origin. Multicollinearity of the independent variables is not a problem as the simple correlation coefficient is .299 for Group I and .166 for Group II.

As the price discrimination model predicts, both cost of hauling and commodity price are highly significant (at the one percent level) in explaining the class rate structure. The regression coefficient for the cost of hauling is positive and greater than unity, and the coefficient of commodity price is positive but less than unity as is also predicted by the model. The unconstrained linear regression explains 70.5 percent of the variance in the rate structure for Group I commodities; for Group II commodities, the amount of variance explained is 86.5 percent. However, an examination of the residuals indicates, particularly for Group I, that the equations consistently overestimate the actual rates for the high cost (less-than-truckload, 260 miles) shipments and underestimate the actual rates for low cost (truckload, 100 miles) shipments. Also, using the

<sup>12</sup> If price  $p=c+r$ , then  $r=[\alpha/(\alpha-1)]K+(p-r)/(\alpha-1)$  or  $r=K+(1/\alpha)p$ . However, this is only an approximation as the price includes the average effect of the transport rates.

<sup>13</sup> A monopoly maximizes profits when  $p=[\alpha/(\alpha-1)](c+r)$ . Therefore  $r=[\alpha/(\alpha-1)]K+[1/\alpha-1][p(\alpha-1)/\alpha-r]$  becomes  $r=K+[(\alpha-1)/\alpha^2]p$ . Again this is approximate given that only data on average price including average transport rate is available.

TABLE 1—REGRESSION EQUATIONS<sup>a, b</sup>

Group I Commodities			R <sup>2</sup>
(1) Rate =	-15.0658 + 1.3393 Cost (0.0691)	+ 0.001943 Price (0.000208)	.706
(2) Rate =	22.6601 + 1.00225 (Cost + Price) (5.8332) (0.00021)	— Price	.679
or =	22.6601 + 1.00225 Cost	+ 0.00225 Price	
(3) Rate =	1.00268 (Cost + Price) (0.00018)	— Price	.661
or =	1.00268 Cost	+ 0.00268 Price	
Group II Commodities			
(1) Rate =	15.2711 + 1.0647 Cost (0.0311)	+ 0.002392 Price (0.000147)	.865
(2) Rate =	22.3047 + 1.00244 (Cost + Price) (2.1169) (0.00015)	— Price	.863
or =	22.3047 + 1.00244 Cost	+ 0.00244 Price	
(3) Rate =	1.00317 (Cost + Price) (0.00015)	— Price	.804
or =	1.00317 Cost	+ 0.00317 Price	

<sup>a</sup> The standard error of the coefficient is shown in parenthesis.

<sup>b</sup> All variables are measured in cents per cwt. and there are 264 observations.

two coefficients to estimate the average value of  $\alpha$  gives inconsistent results.

Constraining the regression equation so that the commodity price coefficient equals the cost of hauling coefficient less unity reduces the multiple correlation coefficient only slightly and eliminates the systematic variation in the residuals with the cost of hauling. However, the computed rates exceed the actual rates for 67 percent of the rates in Group I and 57 percent in Group II.

If the regression equations are further constrained to eliminate the constant, the reduction in the multiple correlation coefficient is again slight. Now, though, the residuals show that the regression equations underestimate the actual rates in 72 percent of the cases of Group I and 77 percent for Group II.

In addition to predicting a strong positive correlation between the variables, it is argued above that the regression coefficients can be used to predict the average value of

the elasticity of demand. Here the results are uncertain. All the constrained consistent coefficients lead to unrealistically high estimates of the average value of  $\alpha$  (in the hundreds). As mentioned above there are at least two reasons why the regression coefficients are likely to overestimate  $\alpha$ —the effect of transport rates on the price of the product and the effect of monopoly pricing by the shipper. With both these effects the regression coefficient for the cost of hauling approaches unity and the price coefficient becomes smaller, particularly when monopoly pricing by the shipper is involved. Even adjusting for these biases, however, the estimated elasticity of demand is still too high.

An explanation perhaps can be found in the ICC requirement that class rates for given weights have the same percentage relationship among themselves regardless of distance. This may not allow the variation in rates required for full profit maximization. Some rates are probably too high and others

too low. The second set of regressions, which gives the best results in terms of the residuals, has large significant constants. If more variation in class rates were allowed the constants would probably be smaller and the regression coefficients larger, yielding smaller estimates of the elasticity of demand.

If this is the explanation, one is forced to conclude that although average rates are probably as high as predicted by the price discrimination model, variation in rates is not large enough for maximum profits.

### V. Conclusions

Although estimates of the average product elasticity of demand are poor, the 1966 Middle Atlantic Conference class rates do seem to be explained by price discrimination. It appears that motor common carriers not only still practice value-of-service pricing but they do so in a surprisingly systematic manner. Cost of hauling and commodity price together explain the class rate structure much better than either variable alone can do.<sup>14</sup> If one considers the inadequacy and biases of the data, particularly with regard to adjustment for differences in cost of hauling characteristics, price discrimination may be substantially greater than that shown by the various regression equations. Regulation by the Interstate Commerce Commission has allowed a portion of the potentially competitive motor carrier industry to act as a price discriminating monopoly.

### REFERENCES

- W. Adams, "The Role of Competition in the Regulated Industries," *Amer. Econ. Rev. Proc.*, May 1958, 48, 527-43.
- R. N. Farmer, "The Case for Unregulated Truck Transportation," *J. Farm Econ.*, May 1964, 46, 398-409.
- R. C. Fellmeth, *The Interstate Commerce Commission*, New York 1970.
- T. C. Koopmans, "Optimum Utilization of the Transportation System," in M. Beckmann et al., eds. *Scientific Papers of Tjalling C. Koopmans*, New York 1970.
- W. Miklius and D. B. DeLoach, "A Further Case for Unregulated Truck Transportation," *J. Farm Econ.*, Nov. 1965, 47, 933-47.
- J. C. Nelson, "The Effects of Entry Control in Surface Transport," in *Transportation Economics*, Universities-Nat. Bur. Econ. Res. Conference Series, New York 1965, 381-422.
- , "Effects of Public Regulation on Railroad Performance," *Amer. Econ. Rev. Proc.*, May 1960, 50, 495-505.
- R. A. Nelson, *Motor Freight Transport for New England*, New England Governors' Committee on Public Transportation, report No. 5, Oct. 1956.
- W. Y. Oi and A. P. Hurter, Jr., *Economics of Private Truck Transportation*, Dubuque 1956.
- J. E. Olson, "Discrimination in Motor Carrier Class Rates," unpublished doctoral dissertation, Brown Univ. 1969.
- E. W. Williams, Jr., *The Regulation of Rail-Motor Rate Competition*, New York 1958.
- Interstate Commerce Commission, *Annual Report*, 1967.
- , Bureau of Accounts, *Cost of Transporting Freight by Class I and Class II Motor Common Carriers of General Commodities, By Regions, for the Year 1966*, 1967.
- , *Freight Forwarder Traffic in Official Territory*, I & S 6497, 301 ICC 65 (1957).
- Middle Atlantic Conference, *Class Rates*, Nov. 1965.
- , *Studies of Characteristics of Traffic and Averages*, Continuing Traffic Report No. 10, CTS 1962 and CTS 1963, Oct. 1965.
- National Motor Freight Traffic Association, *National Motor Freight Classification*, Sept. 1966.
- New England Motor Rate Bureau, *Coordinated Freight Classification* No. 14, Boston, Mar. 1968.
- U.S. Bureau of the Census *U.S. Exports, 1967 Annual*, 1968.
- U.S. Bureau of Labor Statistics, *Wholesale Prices and Price Indexes*, 1967.

<sup>14</sup> For Group I commodities,  $R^2$  for costs of hauling and rates alone is .61 and for commodity prices and rates alone, .28. For Group II commodities the figures are .72 and .26, respectively.

# The Dynamics of Firm Behavior Under Alternative Cost Structures

By GEORGE A. HAY\*

A large and growing number of studies attempt to determine the important factors affecting firms' decisions with respect to price, output, and inventories. A striking feature of this literature is the embarrassingly large number of alternative models—all allegedly consistent with the principles of profit maximization—which are used to justify various reduced form or behavioral equations to be estimated with the appropriate firm or industry data.

It is rare, however, that the equations to be estimated are derived rigorously from the underlying model. Because of this, the restrictions placed on the equations to be estimated are often limited at worst to nothing more than specifying which variables should be included in the regression, and at best to fixing the algebraic signs of some of the coefficients. As a consequence, it is frequently difficult or impossible to discriminate among different models involving the same list of variables.

The present paper is concerned with these problems. In it, I derive the profit-maximizing decision rules implied by a variety of alternative cost structures. The goal of the study is to demonstrate that different assumptions about the cost structure, perhaps equally plausible on a priori grounds, imply differences in the corresponding optimizing behavior, and to point out specifically what those differences are. In addition, since differences among the decision rules are interesting not only in themselves (for purposes of regression analysis) but also because of the differences they imply in the response of firms to changes in the environ-

ment, the decision rules corresponding to the various cost structures are subjected to a dynamic analysis in which various patterns of demand are simulated, and the different behavior patterns compared.

The ultimate goal of the study is to lead to models of the firm that have superior explanatory and predictive power to those encountered to date. Even if that promise is not fulfilled, however, the study should lead to an increased understanding of the way various elements of the objective function interact to generate optimal decisions for the important variables over which the firm has control.

## I

In an early and pathbreaking attempt to determine optimal behavior for a firm with a relatively complex cost structure, Charles Holt et al. employed the *z*-transform approach to generate cost minimizing linear decision rules for production, finished goods inventories, and the size of the work force, for a manufacturing firm whose cost structure could be approximated by a quadratic function. Their work was extended by Gerald Childs who stressed the separate treatment of inventories and unfilled orders. In my earlier article in this *Review*, I included price as a decision variable, thus converting the problem into one of profit maximization. In addition both Childs and I used the results of the model to derive regression equations for use with the appropriate data.

Within the general context of the Holt-Childs-Hay model, there are a number of alternative specifications possible. Although each of the authors justifies the particular specification on a priori grounds, the case is not so strong as to rule out the possibility that an alternative specification might possess improved explanatory and predictive

\* Assistant professor of economics, Yale University. I am grateful to Steven Slutsky for research and programming assistance and to Howard Kunreuther and a referee for helpful comments. An earlier version of this paper was presented at The Second World Econometric Congress, Cambridge, England, September 12, 1970.

power. In any event a systematic treatment of the decision rules implied by alternative cost structures may be useful.<sup>1</sup> Moreover, many of the qualitative results should be applicable not only to models within this rather narrow mold but should extend in some degree to other models which deal with approximately the same list of decision variables.

This paper considers four alternative cost structures within the general class of the Holt-Childs-Hay model. I derive the decision rules implied by the various specifications, and compare the dynamic behavior generated in response to changes in the level of demand.<sup>2</sup> The pattern of demand that is given primary attention is one in which the demand curve is assumed to shift upward by 10 percent for a single period and then return to its normal level. Different results will be obtained depending on the extent to which the change is anticipated.

A second pattern which was tried but not reported is one in which the upward shift is permanent. This experiment serves as a check on the optimality of the decision rules, since the correct equilibrium values for the decision variables are easily calculated, and can be compared with the results of the simulation.

## II. Basic Model

The basic model will be presented here without extensive supporting arguments.<sup>3</sup> The decision variables for the firm are assumed to be the rates of production and shipments, the levels of finished goods inventory and unfilled orders, and price. The following symbols will be employed:

$X_t$  ≡ rate of production in period  $t$

$P_t$  ≡ price in period  $t$

$U_t$  ≡ level of unfilled orders (backlog) at the end of period  $t$

$U_t^*$  ≡ desired level of unfilled orders at the end of period  $t$

$H_t$  ≡ level of finished goods inventories at the end of period  $t$

$H_t^*$  ≡ desired level of finished goods inventories at the end of period  $t$

$O_t$  ≡ new orders in period  $t$  (quantity demanded)

$S_t$  ≡ shipments in period  $t$

$V_t$  ≡ direct unit input costs (labor, raw materials, capital rental) in period  $t$

The following elements make up the objective function:

i) It is assumed that there exists a desired level of unfilled orders for the period  $t$  which is proportional<sup>4</sup> to production during the period. Furthermore it is assumed that the firm incurs a cost for deviating from the desired level which can be expressed as a quadratic. Thus:

$$U_t^* = c_{14}X_t$$

$$C_1 = c_1(U_t - c_{14}X_t)^2$$

ii) Similarly it is assumed that there exists a desired level of finished goods inventory for period  $t$  which is proportional to shipments during the period, and a quadratic cost for deviating from the desired level. Thus:

$$H_t^* = c_{24}S_t$$

$$C_2 = c_2(H_t - c_{24}S_t)^2$$

iii) It is assumed that the firm incurs a cost for changing the rate of production which can be expressed as a quadratic. Thus:

$$C_3 = c_3(X_t - X_{t-1})^2$$

iv) It is assumed that a similar cost exists for changing prices:

$$C_4 = c_4[(P_t - V_t) - (P_{t-1} - V_{t-1})]^2$$

where  $V_t$  is a measure of unit input cost.<sup>5</sup>

<sup>4</sup> If an additive constant is included, a constant term is added to the decision rules.

<sup>5</sup> This formulation warrants a word of explanation. To begin with, ignore the  $V_t$  terms. Then if we specify that the demand curve represents the quantity that can be sold at any price when all firms charge the same price, the  $C_4$  term expresses the risk (which may be only subjective, i.e., as perceived by the firm) that if a firm initiates a price rise it might not be followed, and if it initiates a price cut, other firms might retaliate by

<sup>1</sup> A similar approach, within the context of somewhat different models, was attempted by Holt and Franco Modigliani.

<sup>2</sup> Experiments were performed with shifts in the marginal cost curve but these are not reported here.

<sup>3</sup> For additional discussion, see Hay (1970b).

v) It is assumed that quantity demanded (new orders) is a linear function of price:

$$O_t = Q_t - bP_t$$

where  $b$  is the constant slope and  $Q_t$  the quantity intercept which is assumed to shift from period to period.

vi) The variables are constrained by the following identities:

$$O_t - S_t \equiv U_t - U_{t-1}$$

$$X_t - S_t \equiv H_t - H_{t-1}$$

The constraints can be added to the cost and revenue functions along with a factor  $\lambda$  to discount future profits to yield a Lagrangean expression  $L$  to be maximized:

$$\begin{aligned} L = \sum_{t=1}^N \lambda^{t-1} \{ & P_t Q_t - b P_t^2 \\ & - c_1 (U_t - c_{14} X_t)^2 - c_2 (H_t - c_{24} S_t)^2 \\ & - V_t X_t - c_3 (X_t - X_{t-1})^2 \\ & - c_4 [(P_t - V_t) - (P_{t-1} - V_{t-1})]^2 \} \\ & - \delta_t (Q_t - b P_t - S_t - U_t + U_{t-1}) \\ & - \gamma_t (X_t - S_t - H_t + H_{t-1}) \end{aligned}$$

Analytic solution of the problem using the  $z$ -transform technique<sup>6</sup> would be impossible, involving the solution of an eighth degree equation for the roots of the system. However, the model can be solved numerically when particular values for the cost and revenue parameters are used.<sup>7</sup> The values as-

undercutting. By including  $V_t$  we introduce the notion that a change in direct input costs acts as a signal to the industry so that all firms are expected to pass on such changes in higher (or lower) prices. Each individual firm, therefore, tries to avoid price changes which are more or less than the change in direct input costs. As it turns out, the decision rules with and without  $V_t$  are identical except that with  $V_t$  included, we add to the decision rules an infinite series of terms  $V_{t-1}$ ,  $V_t$ ,  $V_{t+1}$ , ..., with more or less exponentially declining coefficients thereafter.

<sup>6</sup> For an extended treatment of the  $z$ -transform method, see Hay and Holt.

<sup>7</sup> The comments of a referee on this issue are worth repeating. "When we deal with simple static theory, optimum conditions can be expressed in terms of the equality of marginal costs and revenues. However, when

sumed are as follows:

$$b = 2.0 \quad \lambda = .99$$

$$c_{14} = c_{24} = 1.2$$

$$c_1 = 6.5 \quad c_2 = 7.5 \quad c_3 = c_4 = 10$$

The scale of the variables can be set arbitrarily and is here chosen so that the average value of price is 100, and the average value of shipments, production and new orders (in real terms) is 100. Given these values,  $b$  is chosen so that demand elasticity is 2.0. (Recall that the demand curve specifies quantity demanded from the firm when all firms charge the same price.) The values of  $Q$  and  $V$  consistent with these specifications are 300 and 50, respectively. The values of  $c_{14}$  and  $c_{24}$  reflect the observed long-run average of the ratio of finished goods plus goods-in-process inventories to shipments and of unfilled orders to production for U.S. manufacturing. The values of  $c_1$ ,  $c_2$ ,  $c_3$ , and  $c_4$  are chosen so that in any month a 10 percent deviation of the variable in question from the desired level will lead to an increase in costs equal to 10 percent of average monthly revenue, except that  $c_1$  and  $c_2$  are made unequal to avoid a problem of indeterminacy in one of the specifications (Case 3 below). The value of  $\lambda$  reflects an assumed annual discount rate of approximately 12 percent. Several different sets of parameters were tried, and the comparisons among the various models to be discussed were not significantly affected. (For further experimentation with different parameter values, see Hay (1970b)).

The decision rules which represent the solution to the optimization problem with the above cost parameters are presented in Table 1.<sup>8</sup> Only the rules for production,

cost structures are complex and dynamic, the optimal strategies may be so complex that they can not even be solved analytically. Numerical calculations may be the only feasible way to explore the properties of the theory and to deduce its properties. Such calculations can help to develop dynamic theory and clarify estimation problems."

<sup>8</sup> Note that the decision rules include future values of  $Q$ . However, on the basis of work by Herbert Simon and Henri Theil (1957) it is known that in cases of

(Continued)

price and inventory are presented since the rules for unfilled orders and shipments can be derived as linear combinations of those three through the constraints. The choice of which three variables to highlight is therefore arbitrary, and in empirical work might be influenced by data availability.<sup>9</sup>

With regard to the decision rules underlying Table 1 we note that each of the equations is dominated to a degree by the lagged value of the dependent variable, with the coefficient of lagged inventory in the inventory equation being the largest of the three. This is notable since there is no cost-of-inventory change in the model. We also note that the coefficient of  $X_{t-1}$  in the production equation is *not* equal to the coefficient of  $P_{t-1}$  in the price equation, even though the two cost-of-change parameters were set equal. (It is however true that, *ceteris paribus*, increasing the cost-of-change parameter increases the coefficient of the lagged dependent variable in the corresponding decision rule.)

One of the important effects of the decision rules is to determine what part of a shift in the demand curve will be absorbed by increasing price, and what part by an increase in output. For a permanent increase in demand it is easily shown that higher prices will absorb half the increase and higher output the rest. The results for a temporary increase are shown in Figure 1a, which traces the response of the firm to a perfectly forecast 10 percent (30 units) increase in the quantity intercept,  $Q_t$  (i.e., a shift in the demand curve so that an additional 30 units are demanded at every price).

In the period of impact, price rises enough

quadratic criterion functions with linear constraints, substituting expected values of future  $Q$  will lead to decisions which maximize expected profits.

<sup>9</sup> Since there are fewer independent cost parameters than decision rule coefficients, there are constraints involving the latter which ideally should be taken into account in any empirical estimation. This raises the two practical problems of discovering the constraints (which are generally non-linear) and using them. One approach to the first problem would be to calculate the decision rules for a great many sets of cost parameters and regress the calculated coefficients on the parameter values. This would require a large number of cases to provide adequate degrees of freedom.

TABLE 1—DECISION RULE COEFFICIENTS:  
ORIGINAL SPECIFICATION

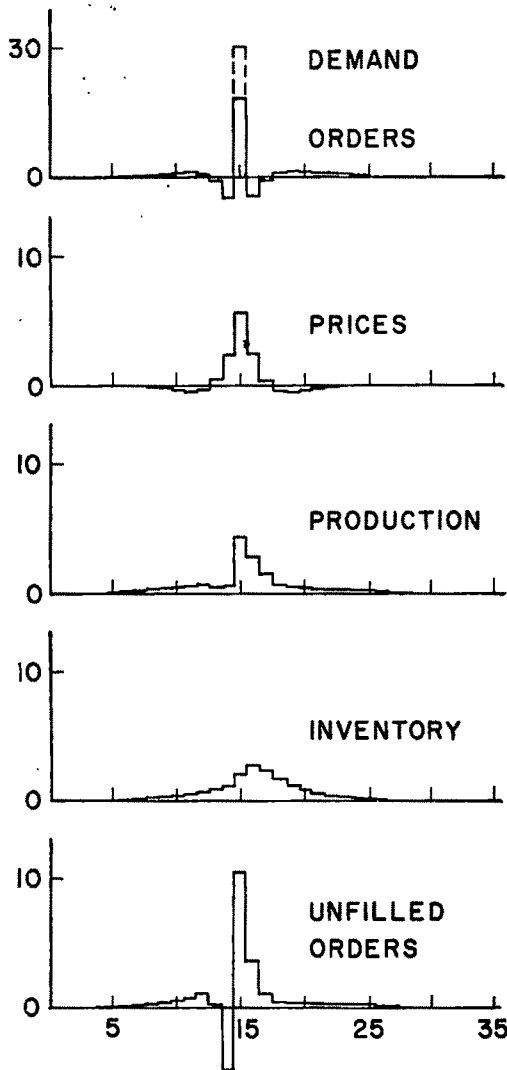
Independent Variable	Dependent Variable		
	$X_t$	$P_t$	$H_t$
$X_{t-1}$	.569	-.302	.364
$P_{t-1}$	-.302	.352	-.125
$H_{t-1}$	-.104	-.100	.504
$U_{t-1}$	.211	.180	.069
$Q_t$	.196	.198	.063
$Q_{t+1}$	.015	.080	.021
$Q_{t+2}$	-.006	.017	.007
$Q_{t+3}^a$	-.001	-.004	.003
$V_{t-1}$	.302	-.352	.125
$V_t$	-.389	.585	-.132
$V_{t+1}$	-.028	-.154	-.042
$V_{t+2}$	.010	-.026	-.013
$V_{t+3}^a$	-.002	.012	-.005

<sup>a</sup> The coefficients of the two infinite series were calculated to period  $(t+15)$  but are not all reprinted here. After one minor oscillation the coefficients decline rapidly toward zero.

to absorb slightly less than 40 percent of the increase, with the rest of the adjustment split between higher production and shipments and an increase in the backlog of unfilled orders. In previous and subsequent periods there is an additional price effect, however, so that the total amount absorbed by price turns out to be half in this case as well. Note that to assist in smoothing output, price actually falls below its long-run level when the demand increase is first anticipated (calculated here to be 15 periods in advance) and falls again after the impact has occurred. Note that while the path of price is symmetric, that of output is not, reflecting the influence of the relationship between  $X_t$  and  $U_t^*$ .

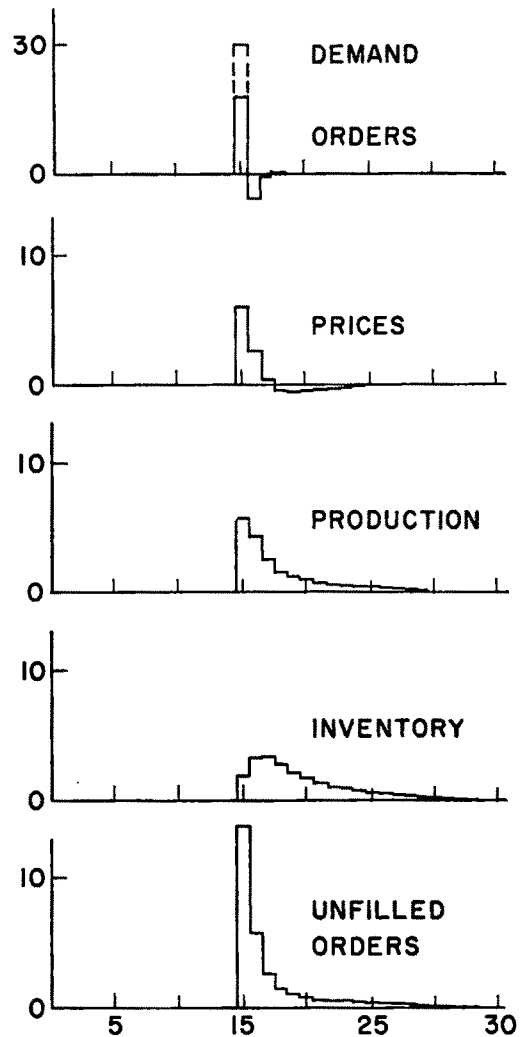
(In Figure 1b we have drawn the case for an increase in demand which is a complete surprise, i.e., not realized until the beginning of the period in which it occurs. Here price still absorbs about 40 percent in the impact period, but subsequent price cutting to smooth the transition to equilibrium output reduces the total long-run effect of price to less than one-third.)

The positive coefficient of  $Q_t$  in the inventory equation suggests that a firm responds



ORIGINAL SPECIFICATION  
(PERFECT FORECAST)

FIGURE 1a



ORIGINAL SPECIFICATION  
(UNANTICIPATED DEMAND)

FIGURE 1b

to an increase in demand by adding to inventories in the period the increase occurs rather than drawing them down. Indeed, both Figures 1a and 1b confirm that observation. The source of this phenomenon is that the increased demand, because it leads to increased shipments in period  $t$ , at the same time raises the desired level of inventories.

The firm must compromise between shipping from inventory and deviating from its desired inventory position, or maintaining its desired inventory position and meeting the extra shipments by sharply raising output. Thus the buffer role of inventories is swamped by the attempt of the firm to maintain its desired inventory position. Some of the spe-

cifications introduced below attempt to modify that result.<sup>10</sup>

We might also note that production, price, shipments and unfilled orders all peak in the impact period, while the peak of inventories lags one period behind. (This result depends however on the parameter values assumed. A lower value of  $c_{24}$  can move the peak of inventories back to the impact period.)

In many similar models a cost of changing inventories has appeared. (See for example Paul Darling and Michael Lovell (1965).) I have argued elsewhere (see Hay (1970a) and the response by Darling and Lovell (1970)), that inventories are primarily a by-product of production. Any lag in bringing inventories to their desired level is due to costs of changing production, since costs specifically applicable to *changes* in the level of inventories are difficult to imagine. To shed more light on this question I added a cost-of-inventory change to the original model, but it turns out that there is very little change in the decision rules (a slightly higher weight on  $H_{t-1}$  in the inventory equation) and virtually none in the response to a one-period change in demand. This result is not particularly surprising since in the original model the conflicting forces operating on inventories produce relatively little movement in that variable; therefore a cost term designed to damp inventory fluctuations should have little effect.

On the other hand, in some of the cases considered below we generate significant movement of inventories and a damping force might be expected to have a more substantial impact. All of the specifications were run with a cost-of-inventory change added with the result that inventory movements became considerably damped (due to a much higher coefficient of  $H_{t-1}$  in the inventory equation) and with production absorbing much more of the adjustment burden for short period shifts in demand.

### III. Alternative Specifications

As mentioned above, the appropriate and exact specification for a model of this type

<sup>10</sup> Unfilled orders are not subject to this cancelling out since the buffer role and the effort to bring unfilled orders to their desired level both act in the same direction.

is difficult to determine on a priori grounds, and several alternative versions can be defended equally well. However, different specifications result in different decision rules and may result in substantially different behavior in response to movements in demand. It is useful therefore to trace through the implications of alternative specifications to highlight their differences. Moreover, the exercise may yield additional insight into the way in which the various parts of the model interact to generate the optimizing decision rules.

#### Case 2

As the first alternative we assume:

$$U_t^* = c_{14}X_{t+1}$$

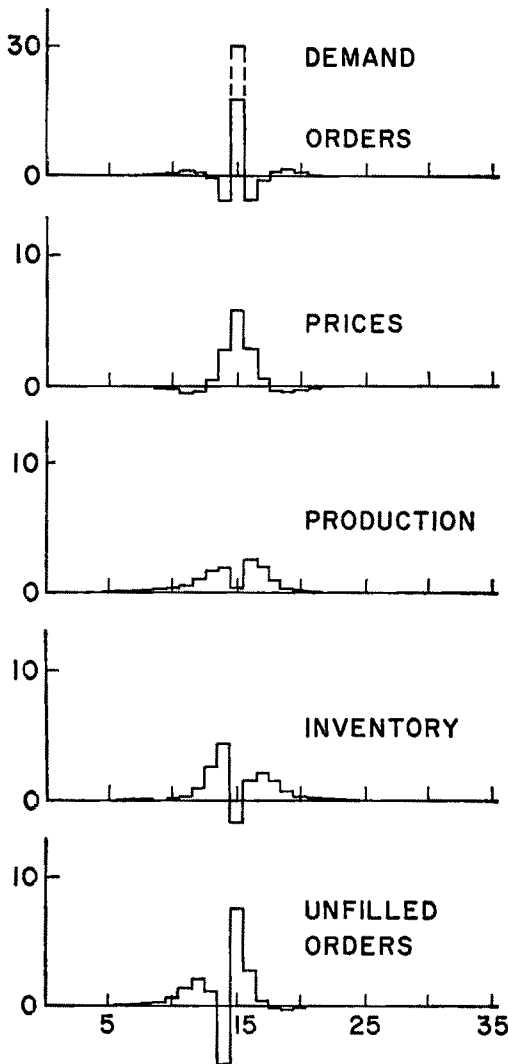
$$H_t^* = c_{24}S_{t+1}$$

The object of this specification is to test the sensitivity of the results to the lag structure. A problem with making the desired end-of-period levels proportional to activity during the period is that inventories tend to overreact to changes in demand, and their role as a buffer stock is effectively cancelled. For this and other reasons it may be interesting to introduce a lead of one period into the desired relationship. (We might note that both versions are observed in the literature, without much attention to the distinction.)

The decision rules corresponding to this specification are presented in Table 2. Sev-

TABLE 2—DECISION RULE COEFFICIENTS  
 $U_t^* = c_{14}X_{t+1}$ ,  $H_t^* = c_{24}S_{t+1}$

Independent Variable	Dependent Variable		
	$X_t$	$P_t$	$H_t$
$X_{t-1}$	.438	— .135	.415
$P_{t-1}$	— .135	.309	— .014
$H_{t-1}$	— .064	— .178	.220
$U_{t-1}$	.412	.092	.242
$Q_t$	.064	.212	— .083
$Q_{t+1}$	.028	.105	.091
$Q_{t+2}$	.012	.023	.044
$Q_{t+3}$	.003	— .012	.009
$V_{t-1}$	.135	— .309	.014
$V_t$	— .137	.552	.146
$V_{t+1}$	— .054	— .202	— .183
$V_{t+2}$	— .021	— .033	— .081
$V_{t+3}$	— .005	.030	— .001



$$U_t^* = c_{14} X_{t+1}, H_t^* = c_{24} S_{t+1}$$

FIGURE 2

eral items deserve comment. First, the production rule is not much changed except that the weight of  $Q_t$  is substantially lower, while the impact of future demand is increased somewhat. In the price rule the coefficient on lagged inventories is larger in absolute value and the weight on unfilled orders smaller. In the inventory decision rule the coefficient of lagged inventories is halved and that of  $Q_t$

actually goes negative while the weight on future  $Q$  is increased.

These changes significantly affect the response of the system to a one period increase in demand, depicted in Figure 2. Since the buffer role of inventories is no longer in conflict with the desired level relation, inventories are built up in the period prior to the demand increase and then are drawn down sharply to absorb about 20 percent of the increase in  $Q$ . As a consequence, unfilled orders and production absorb less of the short-run burden and the buildup of production is more gradual, with the peak coming one period after the demand increase. There is actually a slight dip in the impact period reflecting the influence through the  $U^*$  relationship of the low level of unfilled orders in the previous period.

### Case 3

Here we assume that the desired levels of inventories and unfilled orders are both long-run relationships. Thus the firm would not feel pressure to increase inventories in response to higher demand unless it were thought to be a permanent increase. Similarly, the only short-run upward pressure on unfilled orders is assumed to be the buffer motive. An extreme representation of these assumptions is:

$$U_t^* = \bar{U}, \quad H_t^* = \bar{H}$$

where  $\bar{U}$  and  $\bar{H}$  are treated as constants.

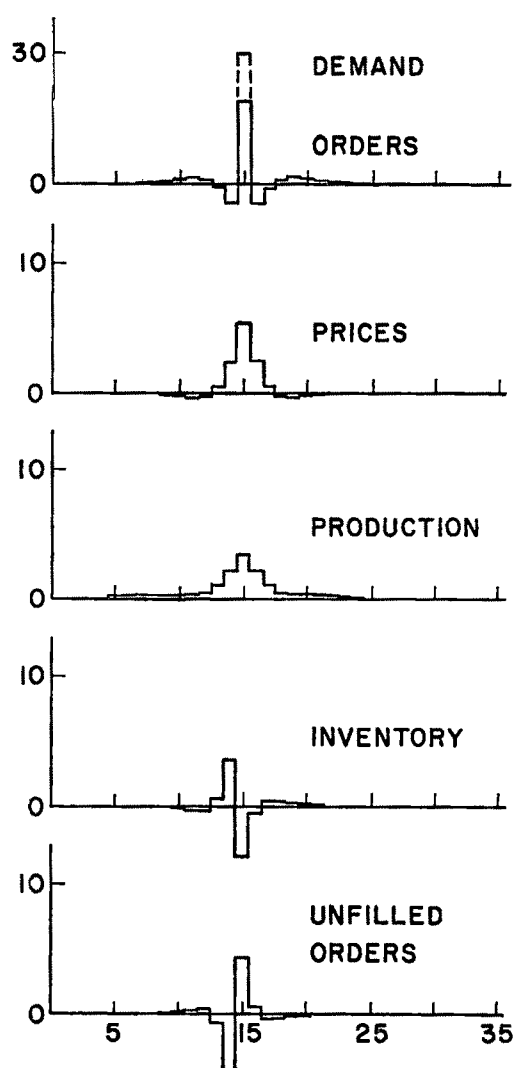
The decision rules derived from this specification are presented in Table 3 and the pattern of dynamic behavior is pictured in Figure 3. We note that the coefficients of  $U_{t-1}$  and  $H_{t-1}$  are equal and opposite in all equations so that under this particular specification it would be possible to treat orders on hand at the beginning of the period as negative inventory and lump the two into a single "net" inventory term. The optimal decisions on the current values of these variables will still be distinct, however, unless we also make  $c_1 = c_2$ . We also note that in the inventory decision, a positive level of unfilled orders at the beginning of the period is now a signal to lower rather than raise the level of inventories. The change results from the breakup of the chain of causation of the

original model in which a backlog of unfilled orders, by leading to more production and therefore higher shipments, caused the desired level of inventories to increase.

TABLE 3—DECISION RULE COEFFICIENTS  
 $U_t^* = \bar{U}$ ,  $H_t^* = \bar{H}$

Independent Variable	Dependent Variable		
	$X_t$	$P_t$	$H_t$
$X_{t-1}$	.698	-.256	.086
$P_{t-1}$	-.256	.303	.163
$H_{t-1}$	-.135	-.210	.207
$U_{t-1}$	.135	.210	-.207
$Q_t$	.122	.225	-.199
$Q_{t+1}$	.056	.092	.111
$Q_{t+2}$	.017	.016	.022
$Q_{t+3}$	.003	-.008	-.006
$V_{t-1}$	.257	-.303	-.163
$V_t$	-.253	.533	.377
$V_{t+1}$	-.109	-.176	-.214
$V_{t+2}$	-.030	-.024	-.036
$V_{t+3}$	-.005	.019	.015
$\bar{H}$	.135	.210	.793
$\bar{U}$	-.135	-.210	.207

We note also that two new variables,  $\bar{H}$  and  $\bar{U}$  are introduced. While the theoretical interpretation of these variables poses no problems, the issue of what to use in regression analysis is not so easily resolved. If we really mean  $\bar{U}$  and  $\bar{H}$  to be constant, they can simply be lumped into the intercept. However the specification does not require that  $\bar{U}$  and  $\bar{H}$  remain constant, but only that the decision maker regard them as being unaffected by his decisions. In the original specification, the decision maker in determining how much to produce had to adjust for the fact that in setting production he was also determining the desired level of unfilled orders, and as we have seen, this feedback effect resulted in a perverse reaction of inventories to a change in demand. For purposes of regression analysis we might still wish to use for  $\bar{U}$  and  $\bar{H}$  some measure of current or anticipated future activity, so that  $\bar{U}$  and  $\bar{H}$  might rise over time, for example, or follow a smoothed out version of the cyclical pattern of demand, while at the same time specifying the model so that the decision maker regards  $\bar{U}$  and  $\bar{H}$  as constant



$$U_t^* = \bar{U}, H_t^* = \bar{H}$$

FIGURE 3

or, at least, completely exogenous. (If the decision maker regards  $\bar{U}$  and  $\bar{H}$  as exogenous but not constant, we would need the series of the expected future values of  $\bar{U}$  and  $\bar{H}$  in the decision rules.) The variables  $\bar{U}$  and  $\bar{H}$  could also be regarded as functions of exogenous variables such as the interest rate.

The nature of the response to a temporary increase in  $Q$  depicted in Figure 3 is similar

to that of the previous case except that the movements of both inventories and unfilled orders now follow a type of symmetry about the original equilibrium. Again however, the buffer roles of those variables show up clearly.

#### Case 4

In the original model, the firm incurs a cost each time the production rate is changed. In particular, if the firm raises output for a single period, it incurs penalties twice—once for raising production and again for returning to the original rate. There are some costs such as hiring and firing costs and setting-up costs which are no doubt directly related to such changes. Other costs, however, such as overtime or idle time might not be adequately captured by such a specification.

As an alternative we might think in terms of a normal rate of production toward which the plant is geared, with a cost of deviating from that rate. In the extreme case, we assume that the normal rate remains constant over time so that the penalty is expressed as:

$$c_3(X_t - \bar{X})^2,$$

where  $\bar{X}$  is regarded as a constant.

The decision rules corresponding to this model are presented in Table 4. The most obvious difference is that  $X_{t-1}$  drops out as an explanatory variable and is replaced by  $\bar{X}$ . This is interesting in view of the fact that in my earlier article in this *Review* as well as in some other studies, lagged production shows surprisingly low (even negative) coefficients, especially when the models are estimated in first differences.

The time paths of production and price (Figure 4) point up clearly the effect of the change. With a temporary rise in  $Q$ , the amount of new demand absorbed by price increase is about double what it was in the original model. Second, although the total amount that has to be absorbed by increased production is correspondingly less, production peaks more sharply than in the original. This reflects the fact that with the present specification, costs can no longer be avoided

by spreading the production rise evenly over several periods.<sup>11</sup>

TABLE 4—DECISION RULE COEFFICIENTS  
 $C_3 = c_3(X_t - \bar{X})^2$

Independent Variable	Dependent Variable		
	$X_t$	$P_t$	$H_t$
$X_{t-1}$			
$P_{t-1}$	-.230	.278	-.087
$H_{t-1}$	-.085	-.122	.514
$U_{t-1}$	.223	.198	.082
$Q_t$	.211	.212	.077
$Q_{t+1}$	-.060	.110	-.027
$Q_{t+2}$	-.019	.035	.001
$Q_{t+3}$	-.002	.003	.004
$V_{t-1}$	.230	-.278	.087
$V_t$	-.429	.560	-.165
$V_{t+1}$	.123	-.214	.054
$V_{t+2}$	.037	-.062	.002
$V_{t+3}$	.005	-.002	-.005
$\bar{X}$	.688	-.433	.423

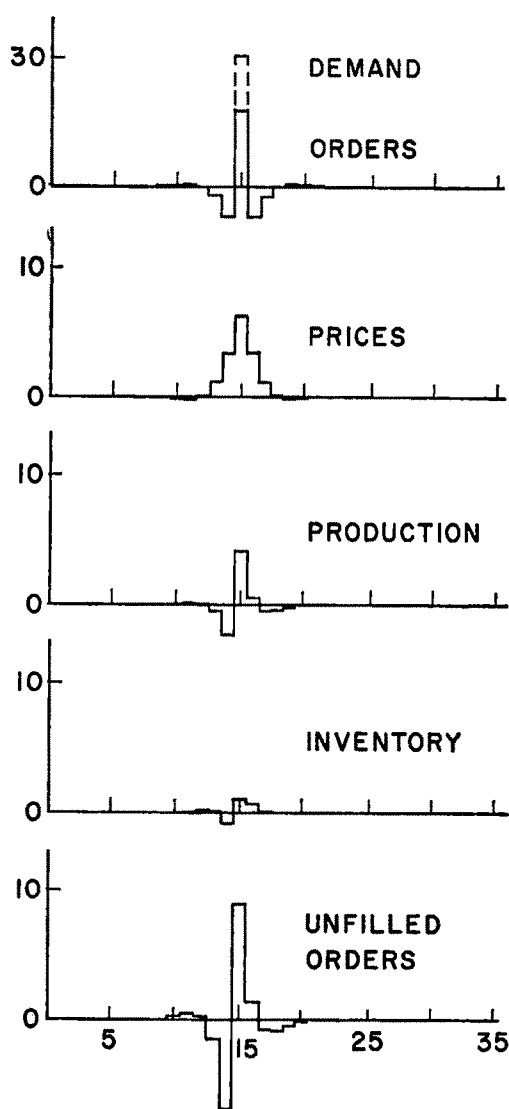
For a permanent increase in  $Q$ , price also absorbs a much larger share since production is tied to  $\bar{X}$  which by assumption does not change. Our remarks concerning  $\bar{U}$  and  $\bar{H}$  are relevant here, since it is likely that the equilibrium level of  $\bar{X}$  would tend to rise if the increase in  $Q$  were indeed permanent.

#### VI. Summary

Obviously a great many specifications are possible even within the narrow context of the Holt-Childs-Hay model and only a few could be presented here. Nevertheless, a number of impressions can be derived from the models tested.

In the first instance, the specification of "desired" levels of inventory and unfilled orders, while a convenience in solving the system, carries certain risks. In particular, when the desired levels are related to the levels of other decision variables in the same period, inventories and unfilled orders tend

<sup>11</sup> In the original model, the firm could absorb increased production requirements of, say, 10 units by spreading it out evenly over 10 periods and incurring costs only for the one unit change in the first and last periods. In the present model, however, a penalty must be paid in every period in which  $X_t$  differs from  $\bar{X}$  so that the disincentive to produce the required amount quickly is reduced.



$$C_3 = c_3(X_t - \bar{X})^2$$

FIGURE 4

to react too strongly to changes in those variables. This produces excessive movements of unfilled orders, where both the buffer motive and the pressure to achieve the desired level operate in the same direction. It also yields perverse movements of inventories where the two motives operate in opposite directions, with the buffer motive swamped by the attempt to maintain desired

inventories. Introducing a one-period lead into the desired relationships allows the buffer role to show up separately. A second possibility is to treat the desired levels as constant at least for moderate time intervals. A third possibility, if a stronger buffer role for inventories is desired, is to substantially reduce  $c_2$ .

Two alternative specifications for the cost of changing production were examined and the behavior of the system was seen to be quite sensitive to the particular specification chosen. The role of price in absorbing demand is twice as important where a constant normal rate of production is specified (with penalties for producing at any other level) compared to the case where only month-to-month changes in production are costly.

A great deal of empirical work has been done in the area of firm behavior regarding prices, output, inventories, etc., and regression results have not always matched perfectly the prior hypotheses. Certainly much of the blame must lie with the data, which are at best imperfect. There is also the problem that models are built at the level of the firm while regressions are run on industry data. It is well known (see Theil (1954)) that only under extremely restrictive assumptions will the industry "decision rule" be a simple scaled-up version of the individual firms' decision rules.<sup>12</sup> The possibility suggested in the present paper is that a survey of the implications of alternative specifications, any one of which can probably be defended on theoretical grounds, may yield some fresh insights into previous empirical work and provide a new basis for planning future efforts.

Although the purpose of the foregoing analysis has been to explore alternative specifications in the specific context of production-inventory type models, it should be stressed that the method of analysis is relevant for a far wider range of dynamic op-

<sup>12</sup> However, as a referee has pointed out, if all the firms in an industry use the same production technology and face the same product and factor markets, then their decision rules should be identical and hence the conditions for perfect linear aggregation would be satisfied.

timization problems. The analysis really applies to any set of cost and constraint structures which meet the mathematical assumptions required for solution,<sup>13</sup> so that its interpretation is by no means limited to production applications. One candidate would be capital theory problems although control theory has proved powerful in that context. Many problems in stabilization policy can also be made to fit the mold, as work by Theil (1964) has shown. Hopefully future work will extend the usefulness of such models even further.

## REFERENCES

- G. L. Childs, *Unfilled Orders and Inventories: A Structural Analysis*, Amsterdam 1967.
- P. Darling and M. C. Lovell, "Factors Influencing Investment in Inventories," in J. S. Duesenberry et al., eds., *The Brookings Quarterly Econometric Model of the United States*, Chicago 1965.
- and ———, "Inventories, Production Smoothing and the Flexible Accelerator," *Quart. J. Econ.*, May 1971, 85, 357–62.
- <sup>13</sup> Specifically, a quadratic criterion function with linear constraints where, if future values of exogenous variables are unknown, the decision maker is content to maximize the expected value of the criterion. This method of solving dynamic optimization problems is given a general treatment in the paper by Hay and Holt.
- G. A. Hay, (1970a), "Inventory Adjustment and the Flexible Accelerator," *Quart. J. Econ.*, Feb. 1970, 334, 140–43.
- , (1970b), "Production, Price and Inventory Theory," *Amer. Econ. Rev.*, Sept. 1970, 60, 531–44.
- and C. C. Holt, "A General Solution for Linear Decision Rules," presented at the European Congress of The Econometric Society, Barcelona, Spain, Sept. 1971.
- C. C. Holt and F. Modigliani, "Firm Cost Structures and the Dynamic Responses of Inventories, Production, Work Force, and Orders to Sales Fluctuations," *Study of Inventory Fluctuations and Economic Stability*, Congress of the U. S. Joint Economic Committee, Dec. 1961.
- , ———, J. F. Muth, and H. A. Simon, *Planning Production, Inventories and Work Force*, Englewood Cliffs 1960.
- H. A. Simon, "Dynamic Programming under Uncertainty with a Quadratic Criterion Function," *Econometrica*, Jan. 1956, 24, 74–81.
- H. Theil, *Linear Aggregation of Economic Relations*, Amsterdam 1954.
- , "A Note on Certainty Equivalence in Dynamic Planning," *Econometrica*, Apr. 1957, 25, 346–49.
- , *Optimal Decision Rules for Government and Industry*, Amsterdam 1964.

# Urban Poverty and Labor Force Participation: Note

By LARRY SAWERS\*

In a recent article in this *Review*, Joseph Mooney examined "the relation between the overall state of the economy . . . and the labor force participation rates of the urban poor" (1967, p. 104). His most interesting conclusion is that the *discouraged worker* effect is much stronger for nonwhite women than it is for white women, and that it is stronger for nonwhite married women than it is for all nonwhite women. In other words, in periods of weak aggregate demand, the labor force participation (*LFP*) of nonwhite women, especially married ones, is depressed more than that of whites. Since this conclusion runs counter to both theoretical considerations and other empirical evidence, the article inspired vigorous dissent.

On theoretical grounds, Jacob Mincer (1962, p. 75) has argued that when the family's principal breadwinner becomes unemployed, the entrance of other family members into the labor force is an alternative to dissaving. Since poor families are unable to dissave as easily as nonpoor, we would expect this *additional worker* effect to be of more importance in poor families. This also means that married women would be expected to have a greater *additional worker* effect than single women; furthermore, since nonwhite families are more likely to be poor than white families, we would expect the *additional worker* effect to be greater in nonwhite families. Mooney's conclusions apparently contradict all of the above hypotheses. Other authors however, have given

some support to Mincer's hypotheses (see Glen Cain (1966, p. 118); Mincer, (1966, p. 95); John Parker and Lois Shaw, pp. 542-44).

Cain and Mincer commented on Mooney's research in this *Review*. The substance of their argument—which is not remarkable for its clarity—is that because Mooney mis-specified his equations, he cannot use the result of his cross-section analysis to make inferences about temporal relationships. In other words, the fact that the *LFP* of nonwhite women relative to white is more influenced by differences in unemployment rates between Standard Metropolitan Statistical Area's (*SMSA*'s) does not necessarily imply that the *LFP* of nonwhite women will be more responsive than that of white women over the business cycle. There is nothing wrong in principle, they argue, with correlating *LFP* with any variable, including those that are indicators of business activity, but before the results can be translated into policy measures, a fuller theoretical understanding of labor supply functions must be derived. Specifically, labor supply is traditionally analyzed in terms of substitution and income effects. Since the *SMSA* unemployment rate is correlated with both wage rates and income levels, the use of the *SMSA* unemployment rate alone as a predictor of *LFP* leaves one hopelessly confused about the true nature of household response to the business cycle. The authors attempt to demonstrate algebraically the conditions under which cross-section and time-series analysis will converge. They then attempt to show through empirical work that Mooney's results are incorrect.

The use of cross-section analysis for projecting onto temporal relationships is a common practice in economic research and the problems associated with the method have long been familiar to economic researchers

\* Assistant professor of economics, The American University. The empirical work in this article was partially supported by Grant No. 91-24-68-75 from the Manpower Administration, U.S. Department of Labor under the authority of Title I of the Manpower Development and Training Act of 1962. Points of view or opinions stated in this document do not necessarily represent the official position or policy of the Department of Labor. Glen Cain and the referee made many helpful comments on an earlier draft of this paper.

(see Edwin Kuh). Two problems are especially relevant for research on *LFP*. First, interurban migration may exaggerate the impact of unemployment on *LFP* if migrants move in response to job opportunities and if they have higher *LFP* than nonmigrants. Empirically, however, this factor seems to have little quantitative importance (see Cain (1966, pp. 68–70, 79); Parker and Shaw, p. 546; and Sawers, p. 68). Second, cross-section unemployment rates almost surely represent long-term structural differences between cities rather than short-term cyclical effects and thus the cross-section response of *LFP* to unemployment may be higher than the time-series one merely because there has been more time for the labor supply to adjust. Thus studies of *LFP* using cross-section data (see William Bowen and Thomas Finegan, Parker and Shaw, Mooney, Sawers) find coefficients on unemployment higher than have been found in comparable analyses of time-series data. The coefficients from cross-section analyses should be thought of as representing longer term effects which are themselves interesting in their own right. One should note that Mooney was well aware of these considerations in his original paper, p. 117.

Cain and Mincer, however, go beyond the preceding discussion and argue that cross-section analysis can be used to make inferences about temporal relationships if substitution and income effects are properly specified. Several comments are in order.

1. In their algebraic exposition of labor supply theory which attempts to show the relation between time-series and cross-section relationships, Cain and Mincer improperly use neoclassical price theory. It is incongruous for them to use a theory which assumes markets are cleared to analyze one which we know is never cleared. (Inflexible wages and lack of mobility and homogeneity are what causes unemployment and these are assumed away in neoclassical theory.) The authors assume that the effects of unemployment on *LFP* works exclusively through the substitution and income effects, both of which are properly thought of as a response to a change in price (the wage rate), not

excess supply (unemployment). There is, of course, a rough correspondence in the real world: wage rates do change somewhat over the cycle and family income changes if unemployment occurs. But viewing unemployment effects only in terms of substitution and income effects is a bit strained. The unemployment rate might have its quantitatively greatest impact on *LFP* if it measures the availability of jobs rather than changes in wage rates or income levels per se and the availability of jobs has little to do with our traditional notions of substitution and income effects.<sup>1</sup> Thus no great faith should be placed in the authors' theoretical discussion.

2. Cain and Mincer's empirical presentation can be flawed on several grounds. First, as Mooney pointed out in his reply, the authors continually identify white with non-poor and nonwhite with poor, even though most of the poor are white and most non-whites are not poor.

The authors' empirical analyses contain both proxies for the income and substitution effects, which are said to represent long-term effects, as well as various measures of unemployment, which are said to represent short-term effects. However, the authors do not mention the intercorrelation between the median income of men (income effect), median earnings of women employed full year (substitution effect), and the various measures of unemployment, and this must surely be very high. To the extent that multicollinearity is present, interpretation of results is difficult. Furthermore, the authors' proxies I would judge to be fairly far removed from what they are trying to measure. Recall that labor supply theory applies to the individual or family while Cain and Mincer use average income in a *SMSA* as a proxy for the income effect. Mooney, of course, uses an *SMSA* variable, the unemployment rate, but he does not try to anchor it so closely to neoclassical price theory.<sup>2</sup>

<sup>1</sup> Mooney in his reply, p. 195, discusses the importance of unemployment as a measure of job availability but does not go on to show how this invalidates Cain and Mincer's theoretical discussion.

<sup>2</sup> An illustration of the difficulties of interpreting this income variable is the fact that other researchers have

Thus by attempting to measure separately the income and substitution effects, the authors have fostered a false sense of precision.

More important, however, Cain and Mincer assert that the income and earnings variables included in their cross-section analysis are "permanent" or "long run" (p. 189) and that the unemployment rates used represent "short-run" or "transitory" effects. The *SMSA* unemployment rate is clearly a long-run variable—in fact, a major criticism of cross-section analysis of *LFP*, as we have noted above, is that the *SMSA* unemployment rate represents a long-term structural characteristic of the *SMSA*. Perhaps the income and earnings variables are "more" long run than the unemployment rate, but this is surely approaching or has passed the stage of quibbling. I suppose—though I am not convinced—that it would be desirable to separate these differential effects, but the data are simply too limited to allow such fine distinctions.

Thus I feel that Mooney was justified in regressing *LFP* on unemployment without other economic independent variables. There are still difficulties in interpreting the coefficient on unemployment, but the superiority of Cain and Mincer's theory and method is not clear. Nevertheless, other important criticisms of Mooney's study could have been made but were not discussed in the published comment.

The most serious objection I might raise to Mooney's research is the inadequate manner in which his data were prepared. He does state that a more extensive preparation of his data "lies beyond the ambitions of this study" (p. 115), but I do not feel that he stressed this caveat sufficiently.

The following criticisms may be taken against Mooney's handling of the data, and

---

tried to use a similar variable as a proxy for the demand for domestics. The two hypotheses predict different signs for the coefficient of income. In addition, the income measures used by Cain and Mincer are not adjusted for variations in the price level. Furthermore, average income levels are surely correlated with the structure of labor demand, and thus the coefficients of the income variables are difficult to interpret. For an analysis of the impact of the structure of labor demand on *LFP*, see Sawers.

all of these criticisms have been dealt with in the empirical analysis presented below. Mooney assumes for statistical purposes that everyone who lived in a census tract where the nonwhite population exceeded 50 percent of the total to be nonwhite even though 49.9 percent could have been white. Conversely, everyone who lived in a tract where the proportion of nonwhites was less than their share in the *SMSA* as a whole were white. Thus his "white population" might have had up to 25 or 30 percent nonwhites.

A second major problem with Mooney's use of the census-tract data is his failure to include demographic variables in his equations. There was only one independent variable, the *SMSA* unemployment rate, in most of his regressions even though it is well established that many demographic characteristics are highly correlated with *LFP*. If large differences in educational attainment or in the age distribution exist between *SMSA*, then his estimates of the coefficients on unemployment may be seriously biased.<sup>3</sup> Mooney makes no attempt to control these factors even though the tract data allow him to do so.

A third major problem concerns Mooney's failure to include the nonpoor as a control group. He is, of course, primarily interested in the poor in his study, but it is obvious that large biases may exist in his estimates because of his data or his research design. Since it is likely that biases for the poor and the nonpoor will be of roughly the same magnitude, the *relative* difference between the two groups may be the same in a biased and in an unbiased analysis.

Many smaller but annoying problems can be found with the manner in which Mooney treated his data. None of these problems by themselves would invalidate his study, but all of them together cast doubt on his results.

<sup>3</sup> If one is interested in only the impact of cyclical factors on equality according to race, there is really no need to adjust for the demographic variables. In this case, it makes little difference whether, as unemployment falls, nonwhites increase their labor force participation rates because of their demographic characteristics or because they are nonwhite. The impact on economic equality according to race will be the same in both instances.

First, Mooney's data are aggregated into white and nonwhite categories. But "non-white" is a racially heterogeneous category, about 94 percent black. Especially in the West, a large proportion of nonwhites are Orientals or Indians and thus the meaning of his results is less clear than if he had analyzed only blacks.<sup>4</sup> Second, many census tracts contain large institutional populations with unusual labor supply characteristics; it is a simple matter to omit these from the analysis even though Mooney did not do so. Third, tracts vary widely in population, from less than 100 to over 25,000 persons. By not weighting his regressions by the population of each tract, bias may have been introduced.<sup>5</sup> Fourth, in his regressions Mooney does not control for the presence or absence of children on the labor force participation of married women. And lastly, Mooney has only forty-five tracts each in his white and nonwhite samples; a larger sample would give one more confidence in the results.

Presented below are the results of a statistical analysis on census tract data that attempts to answer essentially the same questions as Mooney's study, but the data are much more thoroughly developed.

The data pertain to a 15 percent sample of tracts (with over-sampling in poor tracts) in twenty-nine large *SMSA*'s which were chosen to represent a geographical coverage of the United States. The labor force participation rates of all men and women over fourteen years old, for married women, and for married women with children under six years old were computed for the black and white populations of each tract.<sup>6</sup> Proxies which rep-

resent educational attainment, school enrollment, marital status, and age were constructed for men and women of each race for every tract.<sup>7</sup> (However, some of the variables could not be separated by sex and others by race.) Regressions were weighted by the population (white or black) of each tract.

Table 1 shows the regression coefficients and *t*-ratios on the *SMSA* unemployment rate for the twelve subgroups analyzed. Each regression also included the demographic variables listed above, but in the interest of brevity, these coefficients are not presented.<sup>8</sup>

Using a more sophisticated approach to the same data source as Mooney, we find that his estimates of the impact of unemployment on *LFP* were biased upward to a considerable extent, except in the case of white women. However, in a general way, the relative difference in the employment effect on *LFP* between the different subgroups is roughly the same in the two studies. The *LFP* of black women is much more sensitive to labor market conditions than for whites, and is more sensitive for married women than for all women.<sup>9</sup> Furthermore, something that Mooney's study could not show is that this is true in both poor and nonpoor areas. It is interesting to note, however, that the em-

the population in group housing quarters exceeded 15 percent were also excluded.

<sup>7</sup> The proxies were measured as follows: Average levels of educational attainment (percent of population over 25 with a high school degree); school enrollment (percent of population over 14 in school); marital status (percent of population over 14 married and living with spouse); and age (percent of population 14 and over between ages 25 and 54). Educational attainment could not be measured separately for men and women; school enrollment could not be measured separately for whites and blacks.

<sup>8</sup> Coefficients on the five independent variables other than the *SMSA* unemployment rate were generally highly significant with *t*-ratios ranging up to 10. The coefficients of determination (*R*<sup>2</sup>) were generally between .30 and .40.

<sup>9</sup> This is true whether one uses the multiple regression coefficient or the elasticity about the mean. Since black women have higher *LFP* rates than white women, the difference between the coefficients is greater between the two groups than the difference in elasticities. For example, the regression coefficient on unemployment is 40 percent higher for black women in poor areas than white, but the elasticity about the mean is only 25 percent higher.

<sup>4</sup> The problem arises because the proportion of the nonwhite population that is black varies widely from tract to tract, at the extreme from 0 to 100 percent black. A constant proportion would not seriously bias the results.

<sup>5</sup> If the labor force behavior of people in large tracts differs from that of people in small tracts, and if one's conclusions are going to be interpreted as relating to the population of tracts, then weighting will reduce the bias.

<sup>6</sup> Census data by tract are published for the total population and for nonwhites. By subtracting, one can obtain data for whites. Tracts where the nonwhite population was not predominantly black (85 percent or less of nonwhite population was not black) were eliminated from the analysis. Tracts where the proportion of

TABLE 1—PARTICIPATION RATES REGRESSED ON *SMSA* UNEMPLOYMENT RATE  
FOR POOR AND NONPOOR TRACTS, FOR WHITES, AND BLACKS<sup>a</sup>  
(regression coefficients and t-ratios)

	White	Black	Mooney <sup>b</sup>	
			White Tracts	Nonwhite Tracts
<b>Poor Tracts<sup>c</sup></b>	<i>N</i> = 802	<i>N</i> = 501	<i>N</i> = 45	<i>N</i> = 45
Men	-.690 (3.92)	-.573 (2.59)	-2.12 (1.75)	-1.27 (2.48)
Women	-1.93 (7.72)	-2.70 (8.25)	-1.54 (1.65)	-3.61 (4.60)
Women: Married Spouse Present	-2.34 (7.96)	-3.02 (7.60)	1.59 (1.75)	-4.52 (5.70)
Women: Married With Children Under Six	-.450 (3.83)	-1.38 (5.00)		
<b>Nonpoor Tracts<sup>c</sup></b>	<i>N</i> = 994	<i>N</i> = 93		
Men	-.285 (3.91)	-.751 (0.96)		
Women	-1.43 (9.80)	-4.06 (5.17)		
Women: Married Spouse Present	-1.68 (9.89)	-4.99 (5.60)		
Women: Married With Children Under Six	-.305 (4.50)	-1.90 (3.72)		

<sup>a</sup> Other independent variables are age, education, marital status (only in regressions for men and all women), school enrollment, and group quarters population. Their coefficients are not included for economy in presentation. Mooney's regressions used only the unemployment rate as the independent variable.

<sup>b</sup> (1967, pp. 109-10).

<sup>c</sup> Poor tracts are those in the bottom quartile of the *SMSA* income distribution. Mooney defined a poor tract as one where the income was  $\frac{2}{3}$  of the *SMSA* median.

ployment effect is larger for whites in poor areas, but larger for blacks in nonpoor areas.

A second independent variable that Mooney used in his equations for women was an index which reflects the industrial structure of the *SMSA* and measures the specific labor demand for women. This same variable was included in the present study, and Mooney's results are largely supported (see Table 2). Bowen and Finegan, pp. 134-38, 142-44, when they first developed this index found that it had a small but significant impact on the *LFP* of women, i.e., cities with a low demand for female labor (e.g., Pittsburgh) showed a low *LFP* rate for women.

Mooney demonstrated that this is only true for black women. White women in poor tracts showed a negative but insignificant response to a high demand for female workers. The present study supports this finding and shows that the index of demand for female workers has no significant relation to the *LFP* of black women living in nonpoor areas and a *negative* and significant relation for white women in nonpoor areas.<sup>10</sup> Thus the original conclusions of Bowen and Finegan are oversimplified. The only groups to be-

<sup>10</sup> These findings are corroborated in Parker and Shaw, pp. 543-44.

have as they predicted are black women living in poor areas. This is quantitatively a fairly important variable: in this sample from the city with the highest demand for female workers (New York) to the city with the lowest (Pittsburgh), this variable accounted for more than a four percentage point difference in the *LFP* rate of black female slum dwellers.

In summary, the broad conclusions of Mooney's research stand after a much more rigorous use of the same data. It is difficult to reconcile these findings with most other research on the subject. We can, however, confidently state that the problem does not lie with unsophisticated data manipulation.

The problem may lie with the use of small area data. Cain and Mincer note that if "better job opportunities induce and enable white families to move out of poverty areas more easily than nonwhite families, as is generally believed, the white population remaining in the poverty areas will show a weaker labor force response across *SMSAs*" (p. 192). In other words, whites whose *LFP* is responsive to labor market conditions move to the suburbs; but blacks who benefit

from a drop in unemployment rates remain behind in the ghetto because of residential segregation. However, this study shows that participation rates of white women in non-slum areas are lower than those in slum areas (30.0 percent vs. 34.9) and their employment effect is lower. This would indicate that these white families were not able to leave the slums because the wife worked. Rather, the white family made it out on the husband's coattails and the wife was able to *decrease* her participation on the average. In other words, we find some evidence of the *additional worker* effect between slum and non-slum areas for white women. On the other hand, the employment effect for black wives is larger in nonslum areas and her *LFP* is higher (45.9 vs. 42.8 percent). This does not indicate a strong *additional worker* effect. Thus the authors' caveat on intraurban migration does not seem to be important. The use of small area data may still cause bias in our results, but the biases are not as clear as Cain and Mincer would have us believe.

One possible explanation for the higher sensitivity of the *LFP* of black women to unemployment is as follows. It might be

TABLE 2—PARTICIPATION RATES REGRESSED ON *SMSA* UNEMPLOYMENT RATE AND INDEX OF DEMAND FOR FEMALE WORKERS FOR WOMEN IN POOR AND NONPOOR TRACTS, FOR WHITES AND BLACKS<sup>a</sup>  
(regression coefficients and t-ratios)

	Poor Tracts		Nonpoor Tracts	
	<i>SMSA</i> Unemploy- ment Rate	Index of Demand Female Workers	<i>SMSA</i> Unemploy- ment Rate	Index of Demand Female Workers
<b>White</b>				
All women	-2.12 (7.33)	-.159 (1.29)	-1.91 (11.10)	-.467 (5.02)
Married women	-2.29 (6.65)	.042 (0.29)	-1.96 (9.70)	-.271 (2.55)
<b>Black</b>				
All women	-2.18 (5.75)	.550 (2.66)	-4.20 (4.11)	-.132 (0.21)
Married women	-2.91 (6.22)	.105 (0.43)	-4.35 (3.78)	.591 (0.89)

<sup>a</sup> Other independent variables are age, education, marital status, school enrollment, and group quarters population. Their coefficients are not included for economy in presentation.

thought that the unemployment rate across SMSAs would have the greatest impact on opportunities for blacks. If so, and if the black unemployment rate measures opportunity for blacks, then one would expect the ratio of black to white unemployment rates to be higher in slack labor markets. Empirically, however, this is not the case in the sample used for the present study. Another explanation might be that if leisure is a normal good, we would expect a tendency for white women to enter the labor force only when forced to do so by economic necessity, i.e., only when their husband's income fell (the *additional worker* effect). Black women, however, are more likely to be in the labor force *all* of the time because their husbands' income is *always* low, even in periods of prosperity. In periods of high unemployment, black women withdraw from the labor force only because they know jobs are unavailable, and their unemployment is "disguised" as nonparticipation. Thus the results that Mooney and I find are at least plausible.

Nevertheless, the argument is far from being settled. Perhaps the greatest problem with both the study presented here, as well as Mooney, Cain and Mincer's research is that they use aggregated data, either at the tract or SMSA level. The proper unit of observation is the family. With new sources of micro data becoming available (such as the "Survey of Economic Opportunity," "Urban Employment Survey," the "Longitudinal Study of Labor Force Behavior," and the "1970 Census of Population"), perhaps the definitive study can now be done.

It is interesting to note that on the basis of this study the group with the greatest employment effect on *LFP* was not among the poor but was among black women living outside of slum areas. Mooney concluded that the "average nonwhite urban family with both husband and wife present attempt to lift itself out of poverty . . . by becoming multiple-earner families" (p. 117). His statement is based on evidence from slum areas, but it is doubly true in nonslum areas. In fact, the only way that most black families can move out of the slum is by putting the

wife to work. The *LFP* of white women is lower outside of slum neighborhoods than inside but is *higher* outside of slums for black women. White families can move out of and remain out of poverty on the strength of only the husband's income since the white man is not faced with overwhelming discrimination in the labor market.<sup>11</sup> In nonslum areas, however, it is normal for married black women to work and they do so whenever job opportunities are available.

Finally, we should note that not only do white and black women respond to the pressures of aggregate demand in distinctly different patterns, but they respond to the *structure* of that demand in different ways. Black women living in poor tracts were the only group to respond positively in a statistically significant way to the specific demand for female workers. All other groups of women did not respond or responded in the "wrong" fashion. This is evidence that blacks and whites are functioning in labor markets that are at least to some extent separate.

#### REFERENCES

- O. Ashenfelter, "Changes in Labor Market Discrimination Over Time," *J. Human Res.*, fall 1970, 5 168-172.
- W. G. Bowen and T. A. Finegan, "Labor Force Participation and Unemployment," in A. Ross, ed., *Employment Policy and the Labor Market*, Berkeley 1964, 115-61.
- G. G. Cain, *Married Women in the Labor Force: An Economic Analysis*, Chicago 1966.
- , "Unemployment and the Labor Force Participation of Secondary Workers," *Ind. Lab. Relat. Rev.*, Jan. 1967, 20, 275-97.
- and J. Mincer, "Urban Poverty and Labor Force Participation: Comment," *Amer. Econ. Rev.*, Mar. 1969, 59, 185-94.
- E. Kuh, "The Validity of Cross-Sectionally Estimated Behavior Equations in Times Series Applications," *Econometrica*, Apr. 1959, 27, 197-214.

<sup>11</sup> It is a debatable point whether or not cyclical expansion has any *permanent* positive impact on the relative economic status of blacks. See Rasmussen and Ashenfelter.

- J. Mincer, "Labor Force Participation of Married Women," in *Aspects of Labor Economics*, Universities-Nat. Bur. Econ. Res., Conference Series, Ann Arbor 1962, 63-105.
- , "Labor Force Participation and Unemployment, A Review of Recent Evidence," in R. A. Gordon and M. S. Gordon, eds., *Prosperity and Unemployment*, New York 1966, 73-112.
- J. D. Mooney, "Urban Poverty and Labor Force Participation," *Amer. Econ. Rev.*, Mar. 1967, 57, 104-19.
- , "Urban Poverty and Labor Force Participation: Reply," *Amer. Econ. Rev.*, Mar. 1969, 59, 194-98.
- J. Parker and L. Shaw, "Labor Force Participation Within Metropolitan Areas," *Southern Econ. J.*, Apr. 1968, 34, 538-47.
- D. Rasmussen, "A Note on the Relative Income of Non-white Men 1948-1964," *Quart. J. Econ.*, Feb. 1970, 84, 168-72.
- L. Sawers, "The Labor Force Participation of the Urban Poor," unpublished doctoral dissertation, Univ. Michigan 1969.
- Office of Economic Opportunity, "Survey of Economic Opportunity," Washington 1966-67.
- U.S. Department of Commerce, Census Bureau, "1970 Census of Population," Washington 1970.
- U.S. Department of Labor, Manpower Administration, "Urban Employment Survey," Washington 1968.
- , "Longitudinal Study of Labor Force Behavior," Washington 1966-71.

# Nordhaus' Theory of Optimal Patent Life: A Geometric Reinterpretation

By F. M. SCHERER\*

For more than a century the patent system has been studied with remarkable care by economists.<sup>1</sup> Yet only recently, in a contribution by William Nordhaus, has formal economic theory been brought to bear successfully on the central policy issue of the patent system—*how much protection* should be accorded inventors and innovators. This article extends Nordhaus' pioneering work and corrects what in some cases is a significant interpretational error. Nordhaus' original presentation was largely algebraic, but certain problems he left unsolved can be tackled more directly through the geometric approach taken here. This mode of attack has the fringe benefit of making what in the original paper was a rather forbidding tangle of mathematical notation more comprehensible intuitively.

## I. The Basic Model

I begin by observing, as Nordhaus did, that inventions and innovations are not free goods. To make and introduce an invention which reduces unit production costs, research and development (*RD*) outlays must be incurred. For any given production task there exists at some moment in time an "invention possibility function" (*IPF*) which relates the percentage unit production cost reduction *B* achieved (the "output" of an inventive effort) to the expenditure on *RD*. The more research input, the greater will be the cost saving, *ceteris paribus*. For mathematical convenience Nordhaus considers only the very simple invention possibility function  $B = \beta RD^\alpha$ , which with  $\alpha < 1$  implies continuously diminishing marginal returns to inventive effort. However, it seems more plausible (and as we shall see, more flexible)

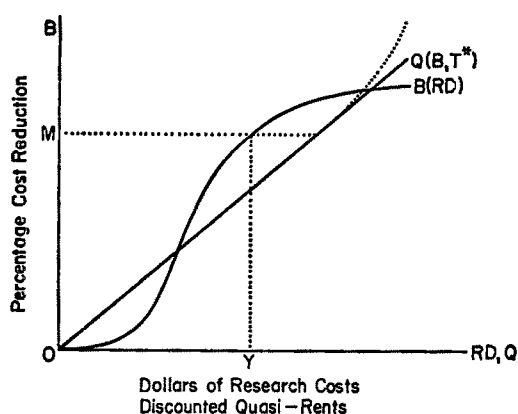


FIGURE 1

to assume an inflected function like  $B(RD)$  in Figure 1, where at first there are increasing returns to the research effort, after which diminishing returns set in.

The benefits to the firm depend in a slightly more complex manner upon the cost savings facilitated. Following Kenneth Arrow, Nordhaus assumes that production is initially carried out under competitive conditions at constant unit cost and price  $OC_0$ , as shown in the conventional supply and demand diagram Figure 2.<sup>2</sup> The firm which secures exclusive patent rights on an invention reducing unit costs to  $OC_1$  can either drive other firms out of business, producing the whole former output  $OX_0$  and commanding a monopoly rent of  $C_0EAC_1$  per year, or it can license the patent to existing producers, charging a royalty which extracts the same surplus  $C_0EAC_1$  from them. Note that even though the patent confers some monopoly power, it does not permit the patent holder to charge a price above the cost  $OC_0$  associ-

\* Professor of economics, University of Michigan. I am grateful to William Nordhaus and the editor for helpful criticisms.

<sup>1</sup> See the survey article by Fritz Machlup.

<sup>2</sup> As long as the new process is merely used internally and not licensed to outsiders, the analysis applies equally well to the case of a monopolist or group of colluding oligopolists pricing to deter all entry in a range of relatively inelastic demand. See Scherer, pp. 219-21.

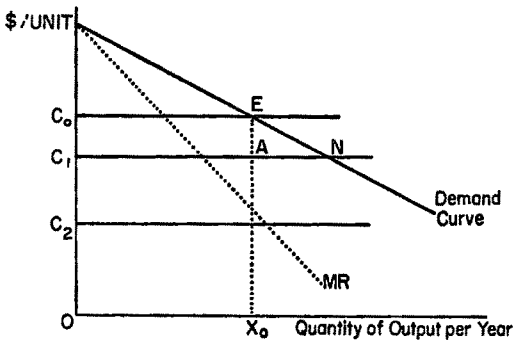


FIGURE 2

ated with the (now inferior) competitive process. Because of this, and if demand is not very elastic in the neighborhood of the competitive price, the optimal postinvention price and quantity under a patent monopoly will be identical to those in the preinvention equilibrium. However, if the invention reduces cost by so much (i.e., to  $OC_2$ ) that the new long-run cost curve cuts the monopolist's marginal revenue curve  $MR$  to the right of the old competitive output  $OX_0$ , the patent holder will find it advantageous to see that price is reduced below  $OC_0$  and that output is expanded beyond  $OX_0$ . Nordhaus assigns the name "run-of-the-mill" to inventions which reduce cost insufficiently to induce a price reduction and output expansion, as with process  $OC_1$  in Figure 2. Following Nordhaus, I shall focus here on the run-of-the-mill case, touching only peripherally on the case of "drastic" price-reducing process inventions like those leading to unit cost  $OC_2$  in Figure 2.

Since there is no output expansion effect with run-of-the-mill inventions, the annual monopoly rent to the patent holder is a linear function of the percentage cost reduction  $B$ . For a given patent life  $T = T^*$ , the patent monopolist's total discounted quasi-rent function  $Q(B, T^*)$  can therefore be shown as a straight line in the coordinates of Figure 1.<sup>3</sup> (If the cost reduction effort is pursued into the stage where the "drastic" invention case holds, the quasi-rent function begins to curve

upward, as shown by the dotted offshoot from  $Q(B, T^*)$  in Figure 1.) Given the market demand and competitive supply conditions, the invention possibility function  $B(RD)$ , and some patent life  $T^*$ , the firm maximizes its profits by extending its  $RD$  expenditures to that level where the horizontal distance between the "cost of invention" function  $B(RD)$  and the quasi-rent function  $Q(B, T^*)$  is a maximum—i.e., at  $RD$  outlay  $OY$  and percentage cost reduction  $OM$  in Figure 1.

The patent grant's duration  $T$  is usually fixed by the government. This means that  $T$  is a parameter to the inventor-innovator (though some exceptions exist). But to the government the patent life is a policy variable—normally decided upon for broad classes of inventions, although case-by-case determination is not inconceivable. As the government increases the patent grant's life under given technological and market conditions, the number of years over which the patent recipient can command monopoly rents rises and the quasi-rent function  $Q(B, T)$  shifts to the right. For simplicity, I (and Nordhaus) assume that the patent holder reaps the full monopoly potential of its invention while the patent is in force and that competitive imitation wipes out supernormal rents completely at the patent's expiration date. Thus, we tentatively ignore imitation lags, reputation effects, and the obsolescence which occurs due to exogenous technological change and competitive "inventing around" one's patent. Given these assumptions and constant market demand, the existence of a positive discount rate ensures that the rightward shift of  $Q(B, T)$  with increases in  $T$  exhibits diminishing returns, since the gains from incremental patent life extensions are discounted more and more heavily. Figure 3 illustrates the relationships between run-of-the-mill invention quasi-rent functions for various patent lives, assuming constant demand and continuous compounding at a 12 percent discount rate. As the period of patent protection rises, the equal-slope, maximum profit point on the invention possibility function shifts to the right, as indicated by arrows designating the sev-

<sup>3</sup> Where  $r$  is the discount rate,  $Q(B, T^*) = B(OC_0) \frac{(1 - e^{-rT^*})}{r}$ . This formulation is used by Nordhaus after setting  $OC_0 = 1$ .

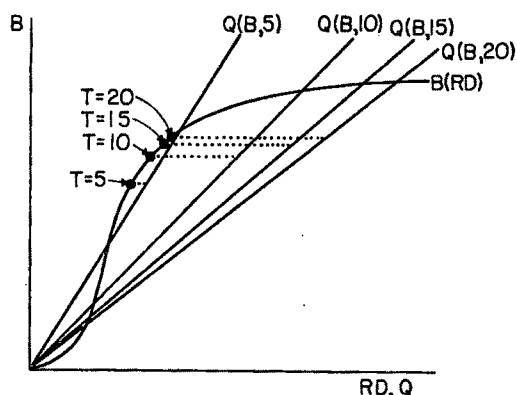


FIGURE 3

eral patent-holder's optima in Figure 3. The longer the patent's life, the further a profit-maximizing firm will carry its cost-reducing  $RD$  expenditures.

How far should this cost-reducing effort proceed if society's interest is to be served? To find out, Nordhaus uses a standard technique of welfare economics. Returning to Figure 2, we observe that while a run-of-the-mill invention leading to unit cost  $0C_1$  enjoys patent protection, the patent owner realizes each year a producer's surplus  $C_0EAC_1$  equal to the cost saving on the preinvention output. When the patent expires, competition drives the price down to  $0C_1$ , output is increased, the producer's surplus is wiped out, and society gains a new consumers' surplus  $C_0ENC_1$ . Ignoring the redistributive implications of this change (or assuming constant and equal marginal utilities of income between the patent holder and society in general), the "price" society pays to induce a reduction in unit costs from  $0C_0$  to  $0C_1$  is essentially the sacrifice of the "welfare triangle"  $EAN (= C_0ENC_1 - C_0EAC_1)$  from the time the invention is introduced until the date of patent expiration, plus the inventor's  $RD$  costs. To find the socially optimal patent life, one must balance the marginal deferrals of this welfare triangle surplus  $EAN$  and the (rising)  $RD$  costs against the increasing amount of cost reduction (and hence increases in producer's and consumers' surplus) stimulated by longer patent lives.

Here geometry proves a blunter tool than

algebra. Nordhaus determines the socially optimal patent life by maximizing with respect to  $T$  the sum of the discounted producer's surpluses  $C_0EAC_1$  from the start of the patent's life ( $t=0$ ) to  $T$ , plus the discounted consumers' surplus  $C_0ENC_1$  from year  $T$  to infinity, less the cost of research and development incurred; all subject to an equality constraint reflecting the inventor-innovator firm's profit-maximizing behavior in setting the slope of its  $IPF$  equal to the slope of its quasi-rent function  $Q(B, T)$ . The mathematics are fairly intricate and will not be reproduced. Here it suffices to characterize the heart of the determination. As the amount of induced cost reduction  $B$  rises due to a longer patent life, society must wait longer and longer to appropriate the welfare triangle  $EAN$ , though with linear demand functions the area of the triangle increases as the square of  $B$ , *ceteris paribus* (see Scherer, p. 402). But for each additional year's wait it motivates less and less incremental cost reduction because the gain to the patent holder increases at best only linearly with increases in  $B$ ,<sup>4</sup> because later years' monopoly rents are discounted more heavily than those in the early years, and because additional percentage points of cost reduction are achieved only at rising research cost. Sooner or later, these diminishing-return effects overpower society's interest in stimulating additional cost reduction by extending the patent life. Therefore, in all but some special limiting cases there exists a finite socially optimal patent life.

## II. Comparative Statics Results

From the constrained welfare-maximizing conditions reflecting this balancing, Nordhaus derives three important comparative statics results. All can be comprehended intuitively from the geometric presentation.

First, the larger is the arc elasticity of demand in the neighborhood of the preinvention and postinvention competitive equilibria, the shorter is the socially optimal patent life. This is so because, at least for run-of-

<sup>4</sup> And in the "drastic" invention case, at less than a linear rate. See the dotted segment of  $Q(B, T^*)$  in Figure 1.

the-mill inventions, nothing varies with elasticity but the base (and hence area) of the welfare triangle  $EAN$ . As price elasticity increases, the area of the welfare triangle increases proportionately, making society less and less willing to postpone its capture in exchange for a given incrementally induced cost reduction.

Second, the "easier" it is to achieve a given cost reduction—that is, the steeper the  $IPF$ , and hence the larger the equilibrium induced level of cost reduction  $B$ , *ceteris paribus*—the shorter the socially optimal patent life will be. This is so because the area of the awaited welfare triangle rises quadratically with increases in  $B$ , as noted earlier, whereas the patent holder's monopoly rent  $C_0EAC_1$  rises only at a linear rate with  $B$ . Hence, when big cost reductions are likely whether the allowed patent life is modest or long, society is less willing to defer the realization of its net welfare surplus to motivate still more cost reduction than it would be if the cost savings under comparable patent life conditions and research investments were modest.

Third, the optimal patent life is shorter, the sharper the curvature of the invention possibility function in the neighborhood of the optimal  $RD$  expenditure. This result is best illustrated with a new diagram. Panels (a), (b), and (c) of Figure 4 each have identical run-of-the-mill invention quasi-rent functions corresponding to patent lives of eight and seventeen years. The invention possibility functions, however, display increasingly

sharp curvatures, analogous to declining elasticities of substitution in orthodox production function theory. The sharper the curvature, the smaller is the difference in the amount of cost reduction induced by a given increase in patent life. And as the cost reduction effect attenuates, *ceteris paribus*, society's welfare gain from deferring competitive imitation falls correspondingly, so shorter patent lives prove optimal.

From Nordhaus' equations 5.13 and 5.14, p. 78, it can be deduced that in the limiting case of a rectangular or "stair-step"  $IPF$ , as illustrated in Figure 4 (c), the socially optimal patent life is zero. Here, however, the theory as developed thus far goes astray by focusing exclusively on the first-order marginal conditions for a patent-holder's local profit maximum, ignoring the necessary *total condition* that expected profits be positive.<sup>5</sup> If the patent's life is set at zero, the quasi-rent function  $Q(B, T)$  coincides with the vertical axis of Figure 4(c), lying at all points but the origin to the left of the  $IPF$ . Faced with a situation in which expected quasi-rents are less than  $RD$  outlays at any positive  $RD$  level, no rational firm will invest, even though the consumers' (and producer's) surpluses potentially available exceed prospective  $RD$  costs. The investment incentive

<sup>5</sup> To deal with this problem algebraically, it is necessary to impose upon the social welfare function being maximized an inequality constraint reflecting the necessity of positive profit expectations, along with Nordhaus' equality constraint reflecting the inventor's marginal optimizing conditions.

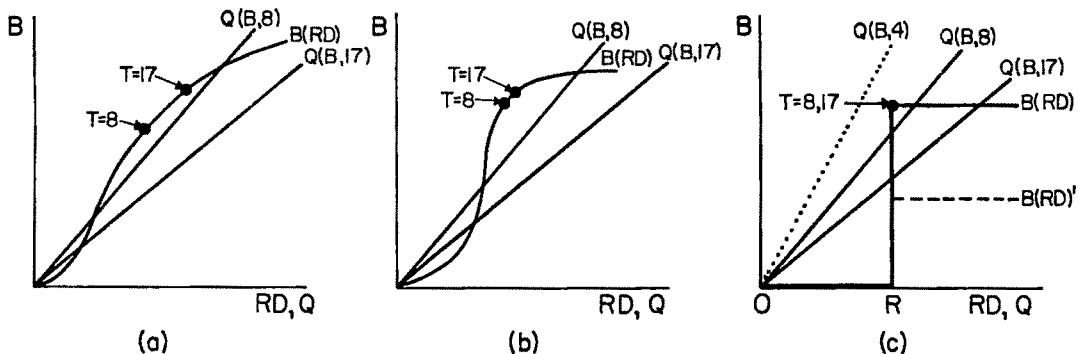


FIGURE 4

is likewise stultified by short patent lives—i.e., with the assumed four-year life indicated by a dotted line in Figure 4(c). And although following Nordhaus' welfare-maximizing marginal rule *necessarily* implies failure to satisfy the positive-profits condition only in this limiting stair-step *IPF* case, the incompatibility can also arise in other cases—notably, when the cost of achieving a given value of  $B$  is relatively high and when the invention possibility function's curvature is relatively sharp in the neighborhood of the inventing firm's equilibrium.

We see then that the patent grant really plays two investment-inducing roles, which in a different context I have called the "stimulus effect" and the "Lebensraum effect" roles (see Scherer, p. 369). In its stimulus effect role, emphasized by Nordhaus, an optimal patent policy sees to it that the monopoly rent lure induces *RD* investment just sufficient to equate the marginal social gain from *further* cost reduction with marginal social cost. In its Lebensraum effect role, the patent grant must persuade investors that competitive imitation will be deferred sufficiently long to make discounted quasi-rents exceed *RD* outlays for at least *some* positive *RD* investment level. Neither aspect can be ignored in designing an optimal patent policy.

Fortuitously, the comparative static relationships derived through Nordhaus' stimulus effect analysis and those associated with the Lebensraum effect are identical for what, from a policy standpoint, is undoubtedly the most important exogenous variable—the "ease" of invention. This is demonstrated in Figure 4(c) by considering two different stair-step invention possibility functions  $B(RD)$  and  $B(RD)'$ . With "hard invention" *IPF*  $B(RD)'$  (broken line), the cost saving realized by spending  $OR$  dollars on research is considerably less than with "easy invention" *IPF*  $B(RD)$ . The Lebensraum condition for  $B(RD)$  is satisfied with a patent life shorter than eight years, but for  $B(RD)'$  even a seventeen-year life is insufficient. Quite generally, the higher is the private benefit/*RD* cost ratio for a potential invention over some assumed period of ex-

ploitation, the shorter the allowed patent life can be while still satisfying the Lebensraum effect condition that expected profits be positive. This conclusion is completely consistent with Nordhaus' finding, considering marginal stimulus effects alone, that easy inventions—those yielding big cost savings in relation to the research resources invested—warrant shorter patent protection than hard inventions.

### III. Some Implications

One immediate policy implication is that if law makers were to shorten the standard term of process patent grants, what would be lost would be primarily those inventions with relatively low benefit-cost ratios—those which in any event are not likely to have a great impact on social welfare. That this conclusion is based upon an extremely naive model does not necessarily nullify its thrust. Indeed, elaboration to account for a principal oversimplification strengthens the case. Contrary to what we have assumed thus far, patent protection is not the only barrier to competitive imitation. Natural inertia, secrecy, and the need to do some *RD* on one's own before mastering a new process all contribute to imitation lags in an atomistically structured industry.<sup>6</sup> When market concentration is high and nonpatent barriers to new entry are present, postinnovation pricing discipline is likely to be sufficiently firm to let an innovating oligopolist recover its *RD* investment even if other immediate rivals imitate quickly. When any or all of these circumstances exist, the Lebensraum condition will tend to be satisfied with short or even zero patent lives for potential process inventions promising high benefit-cost ratios. A uniform policy of long-lived patent grants confers excessive private rewards in these cases, compensated to some unknown extent by the social benefits realized from low benefit-cost projects which otherwise would not have been undertaken and by stimulus effects at the margin of projects which would have been undertaken even with short patent

<sup>6</sup> These points are developed more fully in Scherer, pp. 334-390, which also deals with many of the issues raised in Nordhaus' reply to this paper.

lives. And although the analysis thus far has been developed only for process inventions, it is obvious that for product inventions with high private benefit-cost ratios, the Lebensraum condition will be satisfied at shorter patent lives than for those with low benefit-cost ratios.

A "best of both worlds" policy recognizing these relationships would tailor the life of each patent to the economic characteristics of its underlying invention. This might be achieved through a flexible system of compulsory licensing, under which the patent recipient bears the burden of showing why his patent should not expire or be licensed at modest royalties to all applicants three or five years after its issue. This burden would be sustained if the patentee demonstrated that his invention fell into one or more of the categories in which longer protection is needed to satisfy the Lebensraum condition—e.g., because the market is small relative to the costs of research, or because the cost savings achieved were modest in relation to research costs. When a patent-holding corporation possesses a substantial share of the relevant market and well-established marketing channels, on the other hand, there would be a presumption in favor of early patent licensing or expiration on the assumption that positive innovation profits could normally be attained without the added inducement of strong patent protection.

There is, however, one case in which the implications of the modified Nordhaus theory must be regarded with special caution. Risk and uncertainty—surely significant elements in research and development decision making—have been ignored altogether. For most *RD* projects, Edwin Mansfield's studies suggest, this is probably not a fatal oversight. The uncertainties are sufficiently manageable that they can be taken into account by letting the invention possibility and quasi-rent functions be expressed in terms of expected values or certainty equivalents modified by an appropriate risk premium. Still it is at least conceivable that certain inventions with very high "best guess" benefit-cost ratios require unusually bold, farsighted, time-consuming departures from orthodox

technology, with extraordinary attendant uncertainties and risks. In these cases, strong patent protection offering the prospect of exceptional rewards contingent upon technical and commercial success may be necessary to induce investment. Such cases, which are probably rare (i.e., embracing not more than a dozen or so major inventions per year), cannot safely be squeezed into the mold of an orthodox profit-maximizing model like the one presented here. Nevertheless, the fact that they are exceptions makes it possible to devise a policy dealing with them as such while treating the vast majority of all patented inventions under rules which assume some fairly close approximation to profit-maximizing behavior. In particular, the presumption in favor of early compulsory licensing or short patent lives for inventions with high *ex post* private benefit-cost ratios could be waived upon a showing that the patent recipient exhibited exceptional creativity or undertook unusual technical and/or commercial risks in the invention's development.

Such a policy—flexible, but rooted in the logic of economic theory—would probably be a significant improvement over the present U.S. patent system, recognized by friends and critics alike to be arbitrary and inefficient. It warrants serious consideration.

## REFERENCES

- K. J. Arrow, "Economic Welfare and the Allocation of Resources for Invention," in *The Rate and Direction of Inventive Activity*, Nat. Bur. Econ. Res. conference report, Princeton 1962, 619–25.
- F. Machlup, *An Economic Review of the Patent System*, Study No. 15, U.S. Senate, Committee on the Judiciary, Subcommittee on Patents, Trademarks, and Copyrights, 85th Cong., 2d sess., Washington 1958.
- E. Mansfield, *Industrial Research and Technological Innovation*, New York 1968, ch. 3.
- W. D. Nordhaus, *Invention, Growth, and Welfare; A Theoretical Treatment of Technological Change*, Cambridge, Mass. 1969, ch. 5.
- F. M. Scherer, *Industrial Market Structure and Economic Performance*, Chicago 1970.

# The Optimum Life of a Patent: Reply

By WILLIAM D. NORDHAUS\*

In his geometric discussion of the economics of patents F. M. Scherer has clarified many of the results and pointed to some problems. His discussion is confined to the "pure" theory. I wish in this reply to point out some problems with the pure theory and suggest that the implications are rather different from those drawn in Scherer's article.

## I. The "Pure" Theory of Patents

The pure case of invention, royalty, and patenting in my work and in Scherer's exposition is based on the following assumptions:

1. The supply of inventors (or inventing firms) is perfectly inelastic. Inventors choose the level of inputs to maximize discounted profits.

2. Inventions are "small" (or "run-of-the-mill") process inventions.

3. There is no uncertainty, and the social rate of discount is equal to the private discount rate.

4. Patents confer complete protection over the invention.

5. There is no technological change, cost reduction, or competitive patenting.

6. The product and factor markets are competitive.

Under these assumptions it can be shown that the optimal life of a patent ( $T$ ) is given by the solution to the following pair of equations:<sup>1</sup>

$$(1) \quad \phi = \frac{1 + \eta B}{1 + \eta B(1 + k)}$$

$$(2) \quad B'(R)\phi = rs$$

\* Associate professor of economics, Yale University. Some, but not all, of the comments made by F. M. Scherer on an earlier draft were heeded. Remaining problems are the author's responsibility.

<sup>1</sup> See Nordhaus (1969), equations 5.4 and 5.13.

where

$\eta$  = price elasticity of demand

$R$  = level of inventive inputs

$B = B(R)$  = percentage cost reduction of new process

$r$  = private and social discount rate

$s$  = price of  $R$

$\phi = 1 - \exp(-rT)$

$T$  = optimal life of patent

$k = -B''(R)B(R)/2[B'(R)]^2$

Using the best estimates I could obtain for the parameters, I reached the following conclusions:

First, once a life of six or ten years has been reached, the level of welfare generated by the patent system is very insensitive to the life of the patent.<sup>2</sup>

Second, for small inventions, with percentage cost reduction less than 5 percent, the monopoly losses associated with the patent system are small (less than one-fifth of the gains from invention).<sup>3</sup>

Third, there does not seem to be a strong case for major changes in the life of patents. The only suggestion is that for relatively easy inventions, the life may be too long.<sup>4</sup>

## II. Imperfections in the Market for Information

In my view, the results of Section I are highly suspect unless it can be shown that assumptions 1 to 6 above are either realistic or unimportant simplifications.<sup>5</sup> We now turn to a discussion of these assumptions.

1. The problem of the supply of invention has been raised by Scherer. He has pointed out—quite correctly—that the pure case assumes that profits are positive at the optimal life, which may not be the case. If

<sup>2</sup> See Nordhaus (1969), Tables 5.2 and 5.3.

<sup>3</sup> See Nordhaus (1969), Table 5.4.

<sup>4</sup> See Nordhaus (1969), Table 5.1.

<sup>5</sup> Some of these problems are discussed in Nordhaus (1969) and (1967).

profits are negative, it is natural to assume the inventor will leave the field.

In fact, the analysis of this possibility is straightforward. We can go through the same analysis for the "total" (or Scherer's "Lebensraum") effect as well as the marginal. For this we replace (2) with the profits constraint:

$$(3) \quad V = B(R)\phi - rsR \geq 0$$

where  $V$  is discounted profits. We thus use the "marginal" (or Scherer's stimulus) solution, (1) and (2), if the solution gives  $V \geq 0$ . If the "stimulus" solution gives negative profits, we use the Lebensraum solution:

$$(4) \quad \frac{RB'(R)}{B(R)} \geq 1$$

The condition is that the optimum comes when the elasticity of  $B$  (the "Invention Possibility Function") with respect to  $R$  is greater than unity. This is the positive profit condition. Equation (4) is easily seen to be the implied optimum in Scherer's Figure 4(c) and can be shown rigorously.

The empirical question raised by Scherer is whether using the stimulus solution is likely to give a seriously biased estimate of optimal life. He implies that the rectangular or stair-step *IPF* is more realistic than the smooth substitution case, and we should thus focus on the Lebensraum solution. I have always assumed that diminishing returns hit the inventor with less vengeance than the stair-step case, but, to my knowledge, there is no firm evidence on the degree of curvature of the *IPF*.

2. The pure case refers only to small process inventions. For "drastic" inventions and process inventions,<sup>6</sup> the inventor can recover a smaller fraction of the gains from invention since the royalty is less than the cost reduction. Given other parameters, then, the optimal life is longer for drastic and product inventions.

3. The model outlined above assumes

<sup>6</sup> Drastic inventions are those for which  $B\eta > 1$ , where  $\eta$  is the point elasticity of demand at the original price. Process inventions are those for which there was no output before the invention.

that the inventor's discount rate is equal to the social discount rate. It is more plausible to assume that inventors are risk averse and require a risk premium on invention. Let  $r^*$  be the required rate of return on invention and  $r$  the social discount rate (where  $r^*/r > 1$ ). It can then be shown that the equations above are replaced by

$$(5) \quad \phi = \frac{1 + \eta B}{r/r^* + \eta B(1 + k)}$$

$$(6) \quad B'(R)\phi = rs\left(\frac{r^*}{r}\right)$$

It is easily verified that as  $r^*/r$  rises (that is as the risk discount rises) the optimal life rises while the equilibrium level of invention may go up or down.

4. We have so far assumed that a patent confers complete protection and gives a complete reward to the inventor. A casual glance at the history of technology suggests that many inventions lead to significant further invention which is generally unprotected by the original patent. In the United States, laws of nature are, by statute, excluded from patent coverage. In these examples, the private reward to invention is diminished by the "narrow" coverage of U.S. patent laws.

An important question for patent policy, therefore, is the extent of "breadth" of coverage of patent laws. Let us denote the breadth parameter,  $\theta$ . If the invention lowers cost from  $c_0$  to  $c_1$ , we assume that after the invention the proportion  $1 - \theta$  spills out as freely available technology, giving the competitive cost (that is, the cost of the competitive, freely available process,  $c^*$  as  $c^* = c_1 + \theta(c_0 - c_1)$ ). If there is zero breadth (as in laws of nature), then  $\theta = 0$  and  $c^* = c_1$ , while for complete protection  $\theta = 1$  and  $c^* = c_0$ . Working through our system we find that the optimal life and breadth are given by

$$(7) \quad \theta\phi = \frac{\eta B + 1}{\eta B(1 + k) + 1}$$

$$(8) \quad \theta\phi B'(R) = sr$$

We thus find that life and breadth go hand in hand. The solution to the system is exactly the same for the composite parameter  $\theta\phi$  as it was for the life parameter  $\phi$  when breadth was complete. Thus if breadth is reduced (because giving complete protection is undesirable or impractical) the optimal life must increase to compensate.

5. We have assumed that patenting takes place in a stationary economy, with no general cost reduction or competition from other new inventions. A more realistic assumption would be that the competitive costs decline continuously over the life of the patent. In this case it is clear that the optimal life of the patent is less than the economic life; this allows the government to recoup some of the deadweight loss in the last years of the economic life of the invention. Aside from this point the effect of a progressive economy on patent policy is not clear.

6. The final and perhaps most important problem is the introduction of imperfect competition in the product market. We assume firms try to maximize profits.

In the case of a natural or regulated monopoly—with no threat to entry—patents and the length of patent life are irrelevant. The inventing firm can capture the entire proceeds whether or not a patent has been conferred.

In the more pervasive case of a few large firms and significant barriers to new entry, there is no obvious way of solving the general problem. We can simplify the analysis by assuming that firms do not license inventions to their competitors.<sup>7</sup> In this situation the economic effect of invention will depend on the price response. The inventing firm can keep prices unchanged (as would happen in the run-of-the-mill inventions analyzed above). The social losses and benefits are then exactly those which occur if we treat the firm as an industry. The optimal life is exactly the same as given above. The func-

tion of the patent grant is simply to protect the firm against imitation.

A more likely case is that the firm will lower its price by a fraction of the cost saving in order to increase its market share. Whether it succeeds in increasing its share or not, there will be a gain to the economy during the life of the patent over and above the private gain of the firm. If the firm succeeds in increasing its market share, the resulting output will be produced at lower cost, increasing the average efficiency in the industry. On the other hand, if the competing firms respond to the inventing firm by lowering price, this will squeeze some deadweight monopoly loss out of the system and increase efficiency. In either case, some of the rewards will be passed on to consumers immediately. As in the case of product or drastic inventions, this implies optimal life should be slightly higher for these inventions (other parameters unchanged).<sup>8</sup>

It is puzzling that these conclusions contradict the wisdom of most students of both invention and industrial organization. The reason is probably that many inventions which are patented are so trivial as to be worthy of no patent protection at all. But this is an objection to patent protection for trivial patents, not an argument for less protection for oligopolistic industries.

### III. Conclusions

The theory outlined above is oversimplified but suggestive. Taking account of all the problems, the following conclusions seem to be justified.

First, a fixed patent life is not optimal in theory, although it may be unavoidable in practice. If we are to err on one side, the analysis suggests too long a patent life is better than too short a patent life. For run-of-the-mill inventions, the losses from monopoly are small compared to the gains from invention. The best way to prevent abuse is

<sup>7</sup> There is some fragmentary evidence that licensing of inventions is an unimportant source of revenue from inventions. See Nordhaus (1969), p. 40, fn. 19.

<sup>8</sup> We have assumed that no competitive, or imitative invention occurs. If it does (say because the legal "breadth" of a patent is narrow), then we can apply the conclusions in point 3 above.

to ensure that trivial inventions do not receive patents.

Second, the complications arising from risk, drastic inventions, imperfect product markets, and "inventing around" patents generally point to a longer rather than shorter patent life.

Third, the argument for compulsory licensing without government subsidy is inconsistent with the model of invention used here. Since licensing is feasible in the absence of compulsory licensing, it cannot (in this model) increase the profits from invention and must therefore lower the level of invention. This will be desirable if and only if the

optimal life is less than the actual life (and conversely).

#### REFERENCES

- W. D. Nordhaus, "The Optimal Life of a Patent," Cowles Foundation discussion paper #241, 1967.
- , *Invention, Growth, and Welfare: A Theoretical Treatment of Technological Change*, Cambridge, Mass. 1969.
- F. M. Scherer, "Nordhaus' Theory of Optimal Patent Life: A Geometrical Reinterpretation," *Amer. Econ. Rev.*, June 1972, 62, 428-30.

# A Simple Approach to Existence and Uniqueness of Competitive Equilibria

By DONALD W. KATZNER\*

The questions of existence and uniqueness of equilibria in general, micro-economic models are well known to economists. Since economic reality is thought of only in relation to equilibrium, any meaningful model of the former must permit the latter to exist within it. Assertions concerning uniqueness not only indicate the number of reference points from which reality may be viewed, but also play a substantial role in stability analysis.<sup>1</sup> It is no wonder, then, that existence and uniqueness have been tackled in many alternative contexts and with a variety of analytical techniques.

For the most part, early proofs of existence were essentially based on fixed point theorems. This is true of Arrow and Gerard Debreu, Lionel McKenzie, and Takashi Negishi. David Gale used the lemma of Knaster, Kuratowski, and Mazurkiewicz which implies the existence of a fixed point. In a more restrictive model Henry Wan found an equilibrium price vector by appeal to the separating hyperplane theory. Kiyoshi Kuga was able to prove existence with considerably simpler methods but had to assume that all goods were "weak gross substitutes."

Proofs of uniqueness, although not necessitating as much mathematical sophistication, have in most cases required additional hypotheses. Arrow, Block, and Hurwicz assumed "gross substitutability" while Gale used a weaker version of the same thing. In Wan's approach, uniqueness falls out of strict concavity and smoothness (differentiability) requirements.

The purpose of this paper is to provide a simple approach to existence and uniqueness

which relies on neither fixed point nor separating hyperplane theorems.<sup>2</sup> Although it is not required that all goods be weak gross substitutes, additional hypotheses seem to be the price of mathematical simplification. In what follows these take the form of restrictions on the way price changes affect income distributions.

The technique proposed here involves construction of a general community utility function which generates market demand through maximization subject to an aggregate budget constraint.<sup>3</sup> The additional hypotheses mentioned above ensure that this can be done. Existence of competitive equilibrium then reduces to the question of existence of a maximum of the community utility function subject to an economy-wide transformation surface. The latter issue is resolved by appeal to the mathematical fact that any continuous function defined on a compact set achieves a maximum on that set. Uniqueness of equilibrium prices now depends on smoothness of the transformation or community indifference surfaces, while uniqueness of aggregate outputs rests on strict concavity of the former or strict convexity of the latter. Geometrically the argument justifies use of tangencies between the two to picture competitive equilibria. Existence and uniqueness are thus brought within reach of the beginning student of economics.

<sup>2</sup> This is not to say that fixed points and separating hyperplanes do not exist. They most certainly will. Rather, a proof of existence and uniqueness of equilibrium will be given which is not based on a theorem asserting the existence of fixed points or separating hyperplanes.

<sup>3</sup> Without further assumptions, the community utility function can not say very much about group "utility" or group "welfare." It merely represents a preference ordering defined on the aggregate commodity space from which market demand can be deduced. By the phrase "community indifference surfaces" is meant the level contours of this function.

\* Professor of economics, University of Waterloo. I would like to thank I. F. Pearce for considerable guidance and help with this paper.

<sup>1</sup> See, for example, Kenneth Arrow, H. D. Block and Leonid Hurwicz.

An essential issue which must be overcome to validate the above technique concerns determination of the distribution of income. This is accomplished by assigning to each price vector a unique distribution based on cost minimization by firms. The equilibrium distribution, then, is that corresponding to the equilibrium price vector.

Readers familiar with the literature in this area will note the similarity between the present approach and those of Wan and Negishi. That presented here may be regarded as a generalization of Wan, who is concerned with a more restrictive model in which the distribution problem does not arise. It may also be regarded as a simplification of both since neither the separating hyperplane theorem used by the former nor the Kakutani fixed point and Kuhn-Tucker theorems used by the latter are employed.

## I

The production side of the model under consideration is presented first. Since details and proofs are given elsewhere,<sup>4</sup> only a brief sketch appears below.

Consider a world with  $n$  final outputs produced independently using three inputs: land, labor, and capital. Let the production functions of individual firms be aggregated into industry functions defined for all non-negative input values. Suppose the latter functions are nonnegative, concave (and hence continuous) with strictly convex isoquants, and have positive marginal products at all positive input vectors. Under these conditions, for each vector of aggregate input supplies, a transformation surface,  $T$ , may be defined which shows for any vector of  $n-1$  outputs, the maximum amount of the  $n$ th which can be produced. Furthermore,  $T$  is concave, and if the concavity of industry production functions is strict, then that of  $T$  is also. Note to each point on the transformation surface there corresponds a unique, cost minimizing, vector of inputs for every industry. Hence if output prices are known, input prices may be deduced since they must be equal to val-

ues of industry marginal products. To obtain this last result with constant returns to scale, it is necessary to add the further restriction that output price be equal to average cost.

Denote market quantities of commodity  $i$  by  $X_i$  and prices by  $P_i$ , where  $i=1, \dots, n$ . Write  $X=(X_1, \dots, X_n)$  and  $P=(P_1, \dots, P_n)$ . Then it is easily seen that market output supply functions associate with each  $P$  those outputs,  $X^0$ , which maximize  $P \cdot X$  subject to the transformation surface.<sup>5</sup> Supply functions are single valued only when  $T$  is strictly concave. Under either circumstances their properties are known.

In what follows attention is focused on the special case in which industry production functions exhibit constant returns to scale. The transformation surface, then, may contain flat regions and supply functions could be multi-valued.

## II

On the demand side, let  $k$  run over individuals,  $k=1, \dots, K$ . With  $x_i^k$  denoting quantities of good  $i$  demanded by person  $k$ , and  $M_k$  representing his income, the demand function,  $h^{ik}$ , of this individual is given by

$$x_i^k = h^{ik}(P, M_k)$$

Assume that the  $h^{ik}$  are generated by maximizing well-behaved utility functions subject to appropriate budget constraints, and hence that the  $n-1$  by  $n-1$  matrix of Slutsky functions

$$\|s_{ij}^k(P, M_k)\|$$

is defined, symmetric and negative definite on most of  $\Gamma = \{(P, M_k) : P > 0, M_k > 0\}$  for each  $k$ . Frequently the latter will be true at all points of  $\Gamma$  and it will simplify matters to consider only this case.

Now market demands are sums of individual demands. To obtain them, write

<sup>5</sup> Here

$$P \cdot X = \sum_{i=1}^n P_i X_i$$

<sup>4</sup> See Katzner (1968).

$$\begin{aligned}
 X_i &= \sum_{k=1}^K x_{ik}, \\
 Y &= \sum_{k=1}^K M_k, \\
 d_k &= \frac{M_k}{Y}, \quad k = 1, \dots, K
 \end{aligned}$$

Note  $d = (d_1, \dots, d_K)$  describes the distribution of income among individuals. At the market levels, then,

$$X_i = H^i(P, Y),$$

where the demand function  $H^i$  is defined by

$$\begin{aligned}
 H^i(P, Y) &= \sum_{k=1}^K h^{ik}(P, d_k Y) \\
 (1) \quad &= \sum_{k=1}^K h^{ik}(P, M_k)
 \end{aligned}$$

The income distribution,  $d$ , does not appear as an argument of  $H^i$  since, as indicated later on, it is considered a function of  $P$ .

Market Slutsky functions,  $s_{ij}^H$ , may be stated in terms of Slutsky functions and partial derivatives of individual demand functions. Let  $H_j^i$  be the derivative of  $H^i$  with respect to the  $j$ th price and  $H_Y^i$  that with respect to  $Y$ . Then

$$\begin{aligned}
 s_{ij}^H(P, Y) &= H_j^i(P, Y) \\
 (2) \quad &+ H^i(P, Y) H_Y^i(P, Y),
 \end{aligned}$$

which, from (1), becomes

$$\begin{aligned}
 s_{ij}^H &= \sum_{k=1}^K s_{ij}^k \\
 (3) \quad &+ \sum_{k=1}^K h_M^{ik} \left( \frac{\partial M_k}{\partial P_j} - h^{jk} + \frac{M_k}{Y} \sum_{k=1}^K h^{jk} \right),
 \end{aligned}$$

for  $i, j = 1, \dots, n$ , and where  $h_M^{ik}$  is the derivative of  $h^{ik}$  with respect to  $M_k$  and functional arguments have been dropped to simplify notation. From the symmetry of each  $\|s_{ij}^k\|$  it follows that  $\|s_{ij}^H\|$  will be symmetric if, and only if,<sup>6</sup>

<sup>6</sup> In this case  $\|s_{ij}^H\|$  is also the  $n-1$  by  $n-1$  matrix.

$$\begin{aligned}
 (4) \quad &\sum_{k=1}^K h_M^{ik} \left( \frac{\partial M_k}{\partial P_j} - h^{jk} + \frac{M_k}{Y} \sum_{k=1}^K h^{jk} \right) \\
 &= \sum_{k=1}^K h_M^{jk} \left( \frac{\partial M_k}{\partial P_i} - h^{ik} + \frac{M_k}{Y} \sum_{k=1}^K h^{ik} \right),
 \end{aligned}$$

for  $i, j = 1, \dots, n-1$ . Negative definiteness of  $\|s_{ij}^H\|$  means

$$\begin{aligned}
 (5) \quad &\left\| \sum_{k=1}^K s_{ij}^k \right. \\
 &\left. + \sum_{k=1}^K h_M^{ik} \left( \frac{\partial M_k}{\partial P_j} - h^{jk} + \frac{M_k}{Y} \sum_{k=1}^K h^{jk} \right) \right\|
 \end{aligned}$$

must be negative definite on  $\Gamma = \{(P, Y) : P > 0, Y > 0\}$ .<sup>7</sup> Since the termwise sum of negative semidefinite matrices, at least one of which is negative definite, is also negative definite, a sufficient condition for  $\|s_{ij}^H\|$  to be negative definite is that

$$\left\| \sum_{k=1}^K h_M^{ik} \left( \frac{\partial M_k}{\partial P_j} - h^{jk} + \frac{M_k}{Y} \sum_{k=1}^K h^{jk} \right) \right\|$$

be negative semidefinite. But this is much stronger than is actually required to prove the result. Note that Ivor Pearce's integrability conditions, p. 118, are obtained from (4) by setting

$$\frac{\partial M_k}{\partial P_j} \equiv 0,$$

for all  $k$  and  $j$ .

Once (4) and the negative definiteness of (5) are satisfied, a little more continuity then

<sup>7</sup> Differentiating

$$\sum_{k=1}^K M_k = Y$$

with respect to  $P_j$  and

$$\sum_{i=1}^n P_i h^{ik} = M_k$$

with respect to  $M_k$ , and using these results with (3), it is easily verified that

$$\sum_{i=1}^n P_i s_{ij}^H = 0$$

Therefore the  $n$  by  $n$  matrix of market Slutsky functions can only be negative semidefinite

guarantees the existence of a local community utility function which, when maximized subject to  $P \cdot X \leq Y$ , generates market demand. To obtain a global utility generator still further conditions requiring, in part, boundedness of certain partial derivatives are needed.<sup>8</sup> But these are also relatively minor additions and will be assumed to hold without further discussion. Therefore (4) and the negative definiteness of (5) are sufficient for the existence of a global utility function defined on market aggregates which is twice continuously differentiable, strictly quasi-concave, has positive first-order derivatives, and generates  $H = (H^1, \dots, H^K)$ . The community indifference map consists of the iso-utility surfaces of this function.

Since  $d$  is thought of as a function of  $P$ , the above integration of market demand yields a community utility function which does not depend on the distribution of income. The assumption of a functional relationship between prices and distributions is rather important; distributions must be associated with price vectors in a differentiable manner to validate the argument. A way of making this relationship explicit will now be considered.

Suppose for the moment that the quantities of land and capital available for production are fixed but that the labor supply may vary over all nonnegative values up to some maximum.<sup>9</sup> Then to each aggregate quantity of labor there corresponds a transformation surface as defined in Section I. It is clear that larger labor supplies expand (or at least do not contract) the maximum amount of the  $n$ th commodity which can be produced with the remaining outputs fixed. Assume this relationship between labor supplies and maximum  $X_n$  is continuous. Then there is a continuum of nonintersecting transformation surfaces, each farther from the origin, such that through any  $X$  on or within that surface associated with the maximum labor

supply, a unique transformation surface passes. Denote the "maximal" surface by  $T^0$ .

Let  $(P, Y)$  be given with  $Y$  sufficiently small so that the line  $P \cdot X = Y$ , call it  $L$ , intersects at least one transformation surface below  $T^0$  or  $T^0$  itself. Consider the lowest transformation surface,  $T_L$ , intersecting  $L$ . From Section I,  $T_L \cap L$  consists of the market quantities supplied at prices  $P$  given the fixed resources generating  $T_L$ . If  $T_L$  is strictly concave, then  $T_L$  and  $L$  have exactly one point in common. In such a case input prices are calculated as described in Section I. When  $T_L \cap L$  contains more than one  $X$  (output supply functions are multi-valued) it is assumed that input prices (determined as in the single valued case) are the same for all points in  $T_L \cap L$ . This will be so, for example, when production functions are identical. Thus, regardless of the number of output supply vectors, unique input prices are obtained for each  $P$ .

Now, supposing that each individual contributes known fractions of all land, labor, and capital supplied to producers, the income of every person from these sources can be computed. Since industry production functions exhibit constant returns to scale and since price is assumed equal to average cost, there are no profits to be distributed:  $M_1, \dots, M_K$  and hence  $d$  have been found. For every appropriate  $(P, Y)$  a unique income distribution is therefore obtained. To be consistent with earlier argument it is finally assumed that  $d$  does not depend on  $Y$ . Higher levels of income (i.e., parallel, outward shifts in  $L$ ) result in larger incomes for each person but leave the relative distribution itself unaltered. Thus an individual's share of total income may vary with  $P$  but not with  $Y$ .

Three well-known examples in which community utility functions exist may be viewed as special cases of the above analysis.<sup>10</sup> First

<sup>8</sup> See Katzner (1970) ch. 4.

<sup>9</sup> The labor supply will vary, for example, as labor force participation changes. The more housewives leaving home to work and the more students giving up advanced education for jobs, the larger will be the labor supply.

<sup>10</sup> Actually the theorems quoted from the authors below were originally proved in a different context from that presented above. Their results take the distribution of income as fixed for all  $P$  but apply to all possible distributions. The approach used here permits the income distribution to vary with  $P$  but associates only one distribution to each  $P$ .

(see Paul Samuelson, p. 5n) if all individuals have identical, homothetic utility functions so that their demand functions take the form

$$h^{ik}(P, M_k) = M_k \psi^i(P), \quad k = 1, \dots, K, \\ i = 1, \dots, n,$$

for some functions  $\psi^i$ , then the distribution term in (3),

$$D_{ij} = \sum_{k=1}^K h_M^{ik} \left( \frac{\partial M_k}{\partial P_j} - h^{jk} + \frac{M_k}{Y} \sum_{k=1}^K h^{jk} \right) \\ = \psi^i \sum_{k=1}^K \frac{\partial M_k}{\partial P_j}$$

But since  $\sum_{k=1}^K M_k = Y$ , it follows that  $\sum_{k=1}^K \partial M_k / \partial P_j = 0$ , and hence  $D_{ij}$  vanishes everywhere for all  $i$  and  $j$ . Thus (4) is trivially satisfied and

$$s_{ij}^H = \sum_{k=1}^K s_{ij}^k$$

In the second circumstance (see E. Eisenberg) individuals have homothetic but not necessarily identical utility functions and the distribution of income is not permitted to vary with price changes.<sup>11</sup> Thus

$$h^{ik}(P, M_k) = M_k \psi^{ik}(P),$$

and, since  $Y$  is constant as  $P$  varies,

$$\frac{\partial d_k}{\partial P_j} = \frac{\partial M_k}{\partial P_j} = 0,$$

for all  $i, j$ , and  $k$ . Here  $D_{ij}$  reduces to

$$D_{ij} = - \sum_{k=1}^K M_k \psi^{ik} \psi^{jk} \\ + \frac{1}{Y} \left( \sum_{k=1}^K M_k \psi^{ik} \right) \left( \sum_{k=1}^K M_k \psi^{jk} \right),$$

which is clearly symmetric and satisfies (4).

The third situation (see William Gorman) requires that

<sup>11</sup> This may be interpreted as follows: Start at a tangency between a transformation surface and a line  $L$ . Rotating  $L$  about the point of tangency changes prices while leaving  $Y$  constant. The Eisenberg case requires the income distribution (and hence  $M_k$ ) to remain the same at each new tangency  $L$  achieves with transformation surfaces as it rotates.

$$h^{ik}(P, M_k)$$

$$(6) \quad = f_i^k(P) + \frac{g_i^k(P)}{g^k(P)} [M_k - f^k(P)],$$

where  $f^k$  and  $g^k$  are linearly homogeneous, subscripts on each denote partial derivatives, and

$$(7) \quad \frac{g_i^k(P)}{g^k(P)} = \lambda_i(P),$$

for all  $i$  and  $k$  and some functions  $\lambda_i$ . The demand functions (6) are generated by the indirect utility function

$$v^k(P, M_k) = \frac{1}{g^k(P)} [M_k - f^k(P)],$$

while (7) implies that for each  $P$ , all individuals' Engel curves are straight lines parallel to each other. An easy calculation once again reveals that  $D_{ij} \equiv 0$  and (4) consequently holds.

Actually, the first example above may be regarded as a special case of the third. The property of demand which, for these two situations, permits aggregation is that the income derivatives,  $h_M^{ik}(P, M_k)$ , are independent of  $k$ . Changes in the income distribution due to variations in  $P$  thus have no impact on the accompanying changes in demand. Whenever this condition is met, (4) will be satisfied.

### III

With transformation surfaces and community utility functions in hand the questions of existence and uniqueness of competitive equilibria are easily resolved. Let  $T$  be the relevant transformation surface (recall  $T$  is or lies beneath  $T^0$ ) and  $u^H$  the community's utility function. If  $\bar{X}$  maximizes  $u^H$  subject to the economy lying on or within  $T$ , then taking unity and the  $n-1$  marginal rates of substitution

$$\frac{u_i^H(\bar{X})}{u_n^H(\bar{X})} \quad i = 1, \dots, n-1,$$

as market prices, individuals are maximizing

utility subject to their budget constraints, firms are maximizing profits and quantity supplied equals quantity demanded in all markets. The economy is therefore at competitive equilibrium. Input prices and the distribution of income are determined as indicated earlier. A two dimensional example is pictured in Figure 1.

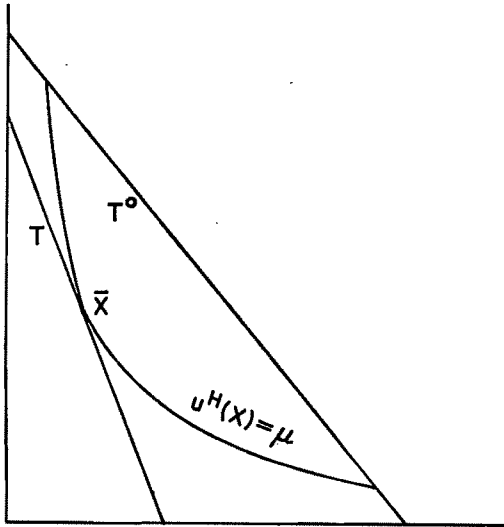


FIGURE 1

Thus existence of equilibrium is reduced to the existence of a constrained maximizer of  $u^H$ . But the collection of all nonnegative points lying on or within  $T$  is a compact set and  $u^H$  is defined and continuous over it. It follows that a constrained maximum and hence competitive equilibrium always exists.<sup>12</sup> Furthermore, equilibrium outputs and normalized prices are unique. These latter assertions are deduced, respectively, from the strict quasiconcavity and positive differentiability of  $u^H$ .

To summarize: within the model de-

scribed here, condition (4) and negative definiteness of (5) guarantee the existence and uniqueness of competitive equilibrium.

# REFERENCES

- T. M. Apostol, *Mathematical Analysis*, Reading, Mass. 1957.
- K. J. Arrow, H. D. Block, and L. Hurwicz, "On The Stability of the Competitive Equilibrium, II," *Econometrica*, Jan. 1959, 27, 82-109.
- K. J. Arrow and G. Debreu, "Existence of an Equilibrium for a Competitive Economy," *Econometrica*, July 1954, 22, 265-90.
- E. Eisenberg, "Aggregation of Utility Functions," *Manage. Sci.*, July 1961, 7, 337-50.
- D. Gale, "The Law of Supply and Demand," *Mathematica Scandinavica*, 1955, 3, 155-69.
- W. M. Gorman, "Community Preference Fields," *Econometrica*, Jan. 1953, 21, 63-80.
- D. W. Katzner, "A General Approach to the Theory of Supply," *Econ. Stud. Quart.*, July 1968, 19, 32-45.
- , *Static Demand Theory*, New York 1970.
- K. Kuga, "Weak Gross Substitutability and the Existence of Competitive Equilibrium," *Econometrica*, July 1965, 33, 593-99.
- L. W. McKenzie, "On the Existence of General Equilibrium for a Competitive Market," *Econometrica*, Jan. 1959, 27, 54-71.
- T. Negishi, "Welfare Economics and Existence of an Equilibrium for a Competitive Economy," *Metroeconomica*, Aug.-Dec. 1960, 12, 92-97.
- I. F. Pearce, *A Contribution to Demand Analysis*, Oxford 1964.
- P. A. Samuelson, "Social Indifference Curves," *Quart. J. Econ.*, Feb. 1956, 70, 1-22.
- H. Y. Wan, Jr., "An Elementary Proof of the Existence and Uniqueness of Competitive Equilibrium in Grahm's Model of World Trade," *Econometrica*, Jan. 1965, 33, 238-40.

<sup>12</sup> A continuous function defined on a compact set always achieves a maximum on that set. See T. Apostol, p. 73.

# Stochastic Dominance vs. Mean-Variance Portfolio Analysis: An Empirical Evaluation

By R. BURR PORTER AND JACK E. GAUMNITZ\*

Most of the work in portfolio theory in the past decade has been based on the principle of utility maximization where either the investor's utility function is assumed to be a second degree polynomial with a positive first derivative and a negative second derivative, or the probability functions are assumed to be normal. If at least one of these conditions holds, it can be shown that choosing among risky assets on the basis of their mean and variance only is consistent with the von Neumann-Morgenstern utility maximization model.<sup>1</sup> Thus, given the above assumptions, the incorporation of higher moments of a distribution and the adoption of alternative approaches to portfolio selection have largely been ignored in favor of the more familiar mean-variance approaches.

After the initial development by Markowitz (1952, 1959), the mean-variance (or *EV* as it is called here) approach to portfolio analysis has been brought to its current state of prominence through numerous extensions and applications such as the Separation Theorem, the simplified or "diagonal" model, the determination of equilibrium asset prices, tests for the efficiency of diversification, and others.<sup>2</sup> Perhaps the most significant

aspect of the *EV* approach has been the concept of *EV* efficiency through which the set of all combinations of risky assets can be separated into two disjoint subsets, one composed of efficient portfolios and another composed of inefficient portfolios. Very simply, a portfolio is efficient if no other portfolio with the same (or smaller) variance has a larger mean and no other portfolio with the same (or larger) mean has a smaller variance. The investor, then, can focus his attention on the efficient subset of portfolios with no loss of expected utility.

Recently, however, an increasing number of writers have challenged the assumptions on which the *EV* approach is founded. Empirical studies have shown that distributions of stock price changes are inconsistent with the assumption of normal probability functions.<sup>3</sup> Moreover, the assumption that utility functions are quadratic in wealth (or return) implies that as risk aversion increases with wealth, beyond some point investors prefer less wealth to more wealth.<sup>4</sup>

In answer to the objections raised by the *EV* approach, a system of preference orderings based on the principles of stochastic dominance was developed by Quirk and Saposnik and extended more recently by Hadar and Russell and G. A. Whitmore. These authors and others have made several important contributions to the theory of portfolio choice.<sup>5</sup> Yet the question remains: Does the application of stochastic dominance (*SD*) rules to portfolio choice yield results that differ significantly from the results that

\* Assistant professors, School of Business, University of Kansas. This study was supported, in part, by the research fund of the School of Business, University of Kansas. We are indebted to James Quirk, Bill Russell, and Bill Breen and the participants in the University of Kansas Finance Workshop for their helpful comments and suggestions. The creative and efficient computer programming necessary for the empirical studies discussed herein was performed by James Wart.

<sup>1</sup> See Harry Markowitz and Marcel Richter.

<sup>2</sup> The list of extensions to the Markowitz model is considerable and not all sources are enumerated here. However, for some of the more important developments the reader is referred to James Tobin, William Sharpe (1963, 1964), John Lintner (1965), John Evans and Steven Archer, Jan Mossin, Michael Jensen, and William Breen.

<sup>3</sup> See Breen and James Savage and Eugene Fama.

<sup>4</sup> For these criticisms and others the reader is referred to William Baumol, Karl Borch, Breen, Josef Hadar and William Russell, Lintner, John Pratt, James Quirk and Rubin Saposnik.

<sup>5</sup> See, for instance, Haim Levy and Giora Hanoch for recent extensions of stochastic models to portfolio choice.

would be obtained using *EV* analysis? This paper presents the results of several empirical studies of the similarities and differences between *EV* and *SD* efficiency.

### 1. The Stochastic Dominance Model

The principles of first, second, and third degree stochastic dominance—*FSD*, *SSD*, and *TSD*, respectively—can be stated essentially as follows:

*FSD*: The probability function  $f(x)$  is said to dominate the probability function  $g(x)$  by *FSD* if, and only if,  $F_1(R) \leq G_1(R)$  for all values of  $R \in [a, b]$  with strict inequality for at least one value of  $R \in [a, b]$ ;

*SSD*: The probability function  $f(x)$  is said to dominate the probability function  $g(x)$  by *SSD* if, and only if,  $F_2(R) \leq G_2(R)$  for all values of  $R \in [a, b]$  with strict inequality for at least one value of  $R \in [a, b]$ ;

*TSD*: The probability function  $f(x)$  is said to dominate the probability function  $g(x)$  by *TSD* if, and only if,  $F_3(R) \leq G_3(R)$  for all values of  $R \in [a, b]$  with strict inequality for at least one value of  $R \in [a, b]$ , and  $F_2(b) \leq G_2(b)$ ; where  $R$  varies continuously on the closed interval  $[a, b]$ ,  $F_n(R) = \int_a^R F_{n-1}(x) dx$ , and  $F_0(R) = f(x)$ .

The significance of these results lies in the fact that each dominance criterion divides the set of all possible combinations of risky assets into an efficient set and an inefficient set—where a portfolio is efficient if, and only if, it is not dominated by another—with the following results:

1. If  $U_1(R) > 0$ ,  $\forall R \in [a, b]$ , then for any portfolio,  $A$ , in the *FSD* inefficient set there exists at least one portfolio,  $B$ , in the *FSD* efficient set such that  $B$  is preferred to  $A$ ;

2. If  $U_1(R) > 0$ , and  $U_2(R) < 0$ ,  $\forall R \in [a, b]$ , then for any portfolio,  $A$ , in the *SSD* inefficient set there exists at least one portfolio,  $B$ , in the *SSD* efficient set such that  $B$  is preferred to  $A$ ;

3. If  $U_1(R) > 0$ ,  $U_2(R) < 0$ , and  $U_3 > 0$ ,  $\forall R \in [a, b]$ , then for any portfolio,  $A$ , in the *TSD* inefficient set there exists at

least one portfolio,  $B$ , in the *TSD* efficient set such that  $B$  is preferred to  $A$ ; where  $U_n(R)$  is the  $n$ th derivative of the investor's utility function.<sup>6</sup>

Since the constraints placed on the utility function by the various dominance criteria are less restrictive than the assumption that utility functions are quadratic, and since stochastic dominance orderings are independent of the type of probability function under examination, the proponents of the *SD* approach to portfolio choice conclude that it is superior to the *EV* approach.<sup>7</sup>

In order to conduct empirical studies of *SD* efficiency we must approximate the true underlying distribution functions by means of finite (and therefore discrete) sets of sample observations. Therefore, the three dominance criteria must be redefined in terms of these discrete observations. This result is obtained by first listing the sample observations in ascending order such that if  $x_i$  and  $x_j$  are the  $i$ th and  $j$ th observations, then  $x_i \leq x_j$  if, and only if,  $i < j$ . Note that although it is possible for two or more observations to have the same numerical value, for consistency in labelling, each observation is considered to be distinct. If there are  $K$  distinct observations of return on a given portfolio, then, each occurs with a relative sample frequency  $f(x_i) = 1/K$ . The corresponding distribution function  $F_1(x_n)$  is generated directly by summing these sample frequencies for all  $x_i, i \leq n$ . Finally, in the comparison of two probability functions,  $f(x)$  and  $g(x)$ , we have a total of  $N = 2K$  distinct observations and the relative frequencies for, say,  $g(x)$  are  $g(x_i) = 1/K$  or  $g(x_i) = 0$ . That is, if the  $i$ th observation belongs to portfolio  $f$  then  $f(x_i) = 1/K$  and  $g(x_i) = 0$ .

<sup>6</sup> For proofs that dominance implies preference the reader is referred to Quirk and Saposnik, Hadar and Russell, Porter, and Whitmore.

<sup>7</sup> *FSD* requires only that the investor's utility function be everywhere increasing and is consistent, therefore, with preference, indifference, or aversion toward risk. *SSD* specifically assumes risk aversion by adding the assumption that  $U_2(R)$  be negative everywhere. *TSD* incorporates the assumptions of *FSD* and *SSD* adding only the requirement that  $U_3(R)$  be everywhere positive.

## II. Empirical Comparisons of *EV* and *SD* Efficiency

In order to gain additional insights into the differences between *EV* and *SD* efficiency, several empirical tests were performed, including: 1) a stochastic dominance analysis of a set of *EV* efficient portfolios; 2) a comparison of the *EV* and *SD* efficient sets obtained from a large set of randomly generated portfolios; and 3) a test of the efficiency of diversification as measured by stochastic dominance as opposed to simple variance reduction.<sup>8</sup>

### *Stochastic Dominance Analysis of the EV Efficient Set*

Monthly data (i.e., monthly rates of return) from 140 stocks<sup>9</sup> for the period 1960-63 were used to generate an *EV* efficient set of portfolios based on the Markowitz Index model formulation. The output from this model gave at each specified value of expected return the corresponding portfolio with the smallest variance. In addition to each specified expected return level, the *EV* efficient set was generated under two different sets of restrictions: 1) where full investment was not required (cash holdings allowed); and 2) where full investment was required with a maximum of 5 percent invested in any one security (cash holdings not allowed). The 5 percent limitation was representative of the normal restriction placed on open-end investment companies and might be considered normal for most other institutional investors.

The means and standard deviations of the six *EV* efficient portfolios generated under the two sets of requirements are presented in

<sup>8</sup> It should be emphasized that all of the tests employed combinations of securities as opposed to individual securities alone.

<sup>9</sup> The stocks were selected in 1966 in connection with a prior study. They were chosen from the 200 largest firms (measured either by total assets or total sales) listed in *Fortune's* 500 as of December 31, 1964. Except for the absence of any firms in the air-transport industry, the 140 selected stocks were representative of the Standard and Poor's 425 industrial stocks. In addition, these stocks were fairly representative of the securities found in the portfolios of many large mutual funds.

TABLE 1—PORTFOLIO SELECTION SUMMARY OF EXPECTED RETURNS AND STANDARD DEVIATIONS 1960-63

	A	B	Portfolio		E	F
			C	D		
I. Full Investment Not Required <sup>a</sup>						
Expected net return for 1960-63 <sup>b</sup>	118.0	30.0	25.0	20.0	10.0	5.0
Standard Deviation	42.0	9.0	7.5	6.0	2.9	1.5
Efficient portfolio held in cash <sup>b</sup>	0	63.6	67.5	73.4	85.6	93.0
II. Full Investment Required <sup>a</sup>						
Expected net return for 1960-63 <sup>b</sup>	118.0	30.0	25.0	20.0	10.0	5.0
Standard Deviation	42.0	11.6	10.4	9.4	7.4	6.4

<sup>a</sup> Maximum of 5 percent in any security.

<sup>b</sup> Percent

Table 1.<sup>10</sup> With one exception, portfolios with identical mean returns showed considerably smaller standard deviations when cash holdings were permitted. This result indicates the variability inherent in portfolios invested 100 percent in common stocks as compared with portfolios having substantial cash holdings. On the other hand, portfolio A, which is the one yielding the highest rate of return (118.0 percent) over the period studied, had no cash holdings. The figures in Table 1 help to illustrate both the classical problem of tradeoff between risk and return and the greater risk-reducing power of additional cash holdings as opposed to additional common stock diversification. One reason for this difference in variance reducing strength is the tendency for returns on common stocks to be positively correlated over time.

The same data from which the means and variances were determined above were used to develop the necessary frequency distributions upon which the *FSD*, *SSD*, and *TSD* tests were applied. In developing these distributions the computationally efficient procedure of classifying observed returns into frequency intervals of equal length was employed. The location of interval end points was adjusted to fit the data, and the number

<sup>10</sup> The returns are reported as total percent return for the holding period rather than as percent per month or per year.

and corresponding length of the intervals were varied experimentally to check for any effect they might have on the results of our efficiency tests. We were able to conclude that the number and length of the intervals did not materially influence our results as long as a reasonable minimum was established along the lines specified in Sturgis' rule (see Ya-Lun Chou).

The results of the stochastic dominance evaluation of the *EV* efficient sets differ drastically depending on whether we are considering the full investment set or the set of portfolios which allow cash holdings. When cash holdings were permitted, the *EV* efficient set was identical with each of the three stochastic dominance efficient sets. The reason for this result is that the portfolios with higher expected returns and greater variability inevitably had at least one outcome smaller than the smallest return observed on any portfolio with a lower expected return. Consequently, no member of the *EV* efficient set was dominated by another when the stochastic dominance criterion was applied.

With full investment required, the results presented in Table 2 are significantly improved. Portfolios *B*, *C*, and *D* were eliminated by *FSD* leaving only *A*, *E*, and *F*—the highest and the two lowest return portfolios—in the efficient set. Portfolio *E* was eliminated by *SSD* and *TSD* leaving *A* and *F* to comprise the efficient set.

Overall, the application of the stochastic

dominance rules is seen to reduce significantly the size of the efficient set as compared with the *EV* rule. This result is consistent with others who have reduced the size of the *EV* set by generally eliminating low return portfolios (see Baumol). However, this test did not determine whether any members of independently derived *SD* efficient sets might be inefficient by the *EV* criterion. Our next comparison overcomes this deficiency.

#### *A Comparison of EV and SD Efficient Sets of Randomly Generated Portfolios*

In order to examine further the relationships between *EV* and *SD* efficient sets, a large number of portfolios were generated using data contained on the Chicago Price Relative tapes.<sup>11</sup> These portfolios were then examined to determine the extent to which the *SD* and *EV* efficient sets conflicted.

Beginning with the 925 stocks with complete data for the 1960–65 period, a number of portfolios containing as few as 2 and as many as 10 stocks each were randomly generated. The resulting portfolios were subsequently recombined, some by random selection and others through a scheme specifically designed to mate portfolios with low correlation. This resulted in increasing numbers of securities in a given portfolio. The process was repeated several times, generating 893 different combinations of individual stocks, before the various efficiency tests were applied.

Of the 893 portfolios ultimately generated, 198 were efficient (not dominated by other portfolios) by *FSD*, 40 by *SSD*, and 31 by *TSD*. The *EV* efficient set contained 39 portfolios, 24 of which were also in the *SSD* set.<sup>12</sup> These results are shown in Figure 1, where an "0" indicates *EV* efficiency, an "x" indicates *SSD* efficiency, and a "□" indicates *TSD* efficiency. The *FSD* results are

TABLE 2—SUMMARY OF PORTFOLIO DOMINANCE  
WITH FULL INVESTMENT REQUIRED  
(1960–63)

Dominating Portfolios	Dominated Portfolios by:	
	First Degree Dominance	Second and Third Degree Dominance
<i>A</i>	<i>B, C, D</i>	<i>B, C, D, E</i>
<i>B</i>	<i>D</i>	<i>C, D</i>
<i>C</i>	<i>D</i>	<i>D</i>
<i>D</i>	None	None
<i>E</i>	None	None
<i>F</i>	None	None

<sup>11</sup> The data referenced were taken from the Price Relative File of the Center for Research in Security Prices at the Graduate School of Business, University of Chicago.

<sup>12</sup> Our results support the findings of Levy and Sarnat derived from their study of efficiency among mutual funds. Of 149 funds in one sample they found 18 to be efficient by *SSD* and 21 by *EV*.

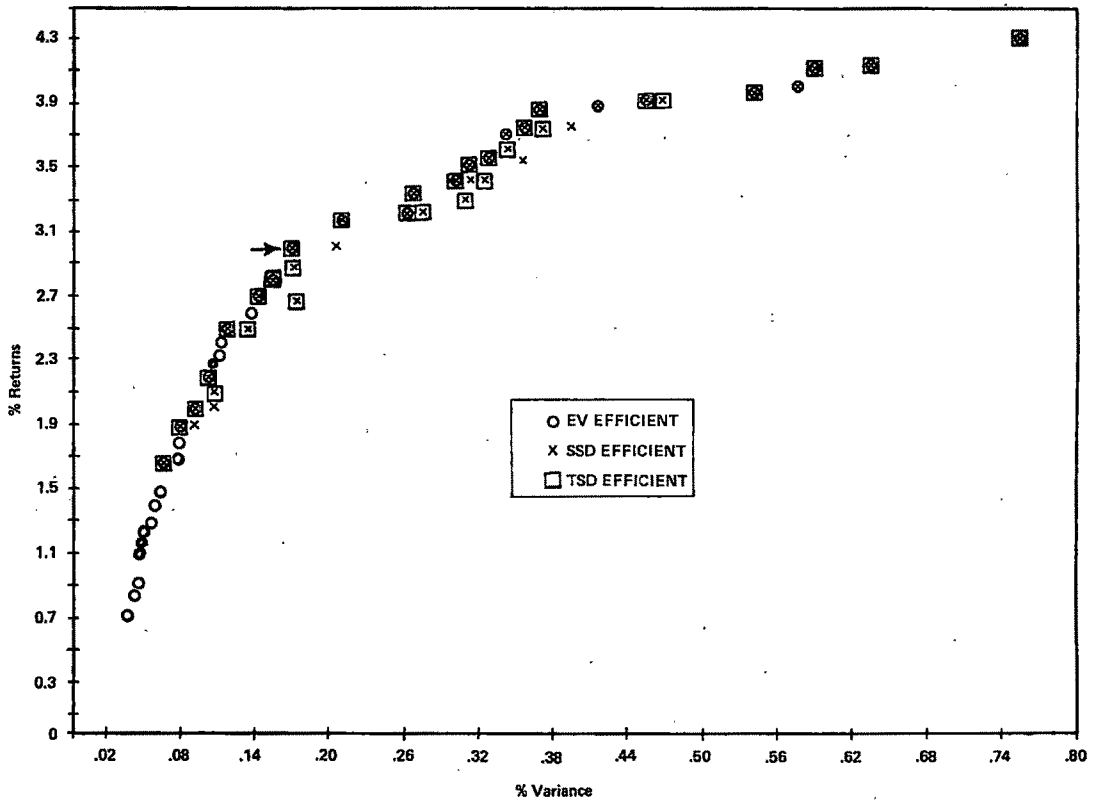


FIGURE 1. *EV*, *SSD*, AND *TSD* EFFICIENT SETS  
(Monthly Rate of Return)

omitted from further consideration because of their lack of significance as compared with the other results. The 15 *EV* efficient portfolios eliminated by *SSD* were relatively low return, low variance portfolios. These findings again verified the conclusions by others on the reduction of the efficient frontier

The 16 *SSD* efficient portfolios eliminated by the *EV* rule were scattered throughout the middle ranges of return and variance, and represented the kind of conflict between the two rules for which we were testing. Further examination of these portfolios revealed that some may not have been eliminated by the *SSD* rule because of gaps in the sample space. For example, in two cases *SSD* efficient portfolios were defeated in the *EV* test by a single *EV* efficient portfolio which almost won by the *SSD* test also. No other portfolio was close enough to present an

effective *SSD* challenge. This conclusion is reinforced by the fact that in no case was an *SSD* efficient portfolio dominated in the *EV* sense by more than two portfolios. In contrast, some *EV* efficient portfolios were dominated in the *SSD* sense by several portfolios.

In addition, several of the *SSD* efficient portfolios eliminated by the *EV* criterion were practically identical to the portfolios that eliminated them, differing only slightly in either mean or variance but having one or two outcomes smaller than the smallest return on the portfolio with respect to which it was *EV* inefficient. For example, one *SSD* efficient portfolio was dominated in the *EV* sense by a single portfolio with the same expected return and a variance that was smaller by only .05 percent. This latter observation deserves emphasis because it suggests a general limitation to the effectiveness

of all stochastic dominance criteria. No matter how high the expected return on a given portfolio and regardless of the other characteristics of the portfolio, if it has one possible outcome that is lower than the lowest possible outcome of another portfolio (regardless of how small the probability of occurrence) the former cannot eliminate the latter using any dominance criteria. In fact, the *SSD* efficient portfolio with the highest expected return failed to eliminate a single *EV* efficient portfolio for precisely this reason. In summary we might note that all of the *SSD* efficient portfolios eliminated by the *EV* rule were relatively close to being *EV* efficient while the converse did not hold. Most of the *EV* efficient portfolios eliminated by either *SSD* or *TSD* were not close to being *SD* efficient. The *SSD* efficient portfolio that was the greatest distance from the *EV* efficient boundary had a mean and variance that differed from the mean and variance of the nearest *EV* efficient portfolio by only .08 percent and .03 percent, respectively. By contrast, the *EV* efficient portfolio that is farthest from the nearest *SSD* efficient portfolio has a mean and variance that differ by .93 percent and .029 percent, respectively. Relatively speaking, then, the greatest deviation from *EV* efficiency within the *SSD* efficient set is much smaller than the greatest deviation from *SSD* efficiency within the *EV* efficient set.

Application of the *TSD* rule eliminated 9 additional portfolios from the *SSD* efficient set leaving a total of 31 efficient portfolios. Of the 9 eliminated, 6 were portfolios that were inefficient by the *EV* rule. Thus, the *TSD* efficient set included only 9 *EV* inefficient portfolios. The *EV* set included 18 *TSD* inefficient portfolios.

Several other results are worthy of note. Of the 925 stocks with which the analysis began only 26 securities are represented in the portfolios which are efficient by *SSD*. The *SSD* efficient set, then, includes a number of differently weighted combinations of these 26 stocks. In addition, the weights seem to vary according to what one would expect, in the sense that high return portfolios included a higher proportion of stocks

that performed well in the period such as Delta Airlines, National Airlines, and Boeing Aircraft, while the low variance portfolios were heavily influenced by the relatively low risk utility stocks such as Washington Gas and Light. Moreover, the high return portfolios included relatively fewer stocks than did the low variance portfolios. The efficient portfolio with the highest mean return was a single stock, and the next two efficient portfolios were two different combinations of two stocks. The largest number of different stocks included in any *SSD* efficient portfolio was 17, which occurred in the *SSD* efficient portfolio with the third smallest variance.

Finally, these same data were subjected to another efficiency criterion test suggested by Levy and Hanoch. This rule is based on the first three moments of the probability function and is so restrictive that only one portfolio can be included in the efficient set unless several have the same first three moments. Our results show that the single portfolio indicated by the arrow in Figure 1 dominated all other portfolios by this test. It is worth pointing out that this portfolio is not the one with the highest expected return, but is located at a point which might be interpreted as that point where the rate of increase in return as variance increases begins to taper off or, alternatively, where the rate at which variance falls as return is reduced also tapers off.

#### *A Stochastic Dominance Test of the Efficiency of Diversification*

One of the applications of *EV* analysis has been the test of the efficiency of diversification where efficiency has been measured in terms of the rate of decrease of the variance as the number of different securities in the portfolio is increased. The results, as reported by Evans and Archer, Lintner, and by Gaumnitz, indicate that most of the gain from diversification is obtained with somewhere between 10 and 20 securities. Since increasing the number of securities in the portfolio generally lowers the expected return, an efficient portfolio should not have more than, say, 12 securities.

Since these tests do not explicitly measure the loss in expected return as variance is reduced, a stochastic dominance test of the efficiency of diversification seems to be more reasonable. To effect this test of efficiency of diversification through *FSD* and *SSD* rules, 7 sets of 20 portfolios each were selected at random from the same 925 stocks discussed above. The first set of 20 portfolios included only 1 stock each, the next set of 20 portfolios included 6 stocks each, then 11, 16, 21, 26, and finally, 31 stocks in each portfolio for a total of 140 portfolios (20 at each grouping). Each portfolio in a given set was compared by *FSD* and *SSD* with each portfolio in every other set. Consequently, at a given level there were 400 possible comparisons. For example, in comparing 6-security portfolios with 11-security portfolios, the distributions of returns for each portfolio in the latter group was compared to the distributions of each of the portfolios in the former group and checked for stochastic dominance. This process was repeated through all portfolios and at all levels and involved a grand total of 6400 different comparisons. If most of the portfolios with a given number of securities dominated those with a smaller number of securities, presumably the additional diversification obtained with the larger number of securities was justified. On the other hand, if the opposite effect occurred, then one might conclude that the practical limit of diversification had been reached.

The *FSD* results were too weak for meaningful analysis in that few portfolios were dominated using the *FSD* criterion. The results of the *SSD* test are presented in Table

3. The first entry for paired values in Table 3 indicates the number of times out of 400 total comparisons that the larger numbered portfolio exhibited second degree dominance over the smaller numbered portfolio listed on the left. For instance, *B* portfolios (6 securities each) exhibited second degree dominance over *A* portfolios on 144 occasions while the opposite result occurred only 31 times. (The remaining portfolio comparisons out of 400 failed to exhibit second degree dominance.)

Looking at the Totals column of Table 3 one can see that most portfolios composed of 6 securities or more generally exhibited second degree dominance over one security portfolios, that is, on 813 occasions larger portfolios dominated 1-stock portfolios and in only 114 cases did the reverse occur. The results of higher level portfolios were inconclusive. The first reversal of second degree dominance in total (last column, Table 3) occurred where portfolios composed of 16 securities dominated portfolios composed of more securities. That is, portfolios with 16 securities each had 408 instances of second degree dominance over larger portfolios, compared with only 252 occasions where the opposite effect obtained. On the other hand, portfolios composed of 21 securities each were generally dominated by larger portfolios (289 versus 168, Table 3). From these results it is difficult to conclude precisely what is the optimum number of securities in a diversified portfolio. Nevertheless, the overall performance of the 16-stock portfolios tends to reinforce the feeling that relatively few securities give ample diversification in common stocks.

TABLE 3—SECOND DEGREE DOMINANCE COMPARISONS OF PORTFOLIOS WITH VARYING NUMBER OF SECURITIES

Portfolio	Number of Securities						Totals
	1 <i>A</i>	6 <i>B</i>	11 <i>C</i>	16 <i>D</i>	21 <i>E</i>	26 <i>F</i>	
<i>A</i>		144, 31	127, 26	150, 11	126, 23	135, 10	813, 114
<i>B</i>			94, 107	127, 56	102, 97	119, 58	552, 394
<i>C</i>				166, 50	127, 111	157, 43	593, 268
<i>D</i>					71, 167	103, 94	252, 408
<i>E</i>						160, 69	289, 168
<i>F</i>							73, 126

### III. Summary and Conclusions

The results of the empirical tests presented in this paper tend to support several conclusions. First, the differences between *EV* and *SSD* efficiency were not as great as might have been expected. The most significant difference was the frequency with which *SD* eliminated *EV* efficient portfolios in the low return range. Clearly, there were other conflicts; some *SSD* efficient portfolios were *EV* inefficient and vice versa. However, except for the disproportionate number of low return portfolios found in the *EV* set, the two sets were very similar. Second, the application of *TSD* reduced both the *EV* and the *SSD* efficient sets and specifically eliminated some of the *SSD* efficient, *EV* inefficient conflicts. Third, *FSD* as a decision criterion was relatively ineffective since few portfolios were eliminated from the efficient set. Fourth, the *SSD* test of the limits of diversification, while not as convincing as the *EV* test, does tend to support the general conclusion that a relatively few securities give adequate diversification in common stock portfolios.

Thus, in general, the most significant difference between the *EV* results and the *SSD* and *TSD* results is the tendency of stochastic dominance to eliminate from consideration the low return, low variance portfolios. Surprisingly, we feel that this conclusion implies that the less risk averse an investor is, the more indifferent he should be with regard to a choice between *EV* and *SSD* as efficiency criteria. The implication follows from the fact that as the degree of risk aversion decreases the individual moves farther up the range of expected value, and the farther he goes in this direction the more similar the *EV* and *SSD* efficient sets become. By contrast, the highly risk-averse investor is the one most likely to suffer from the use of a mean-variance model. The fact that *EV* efficient portfolios in the low return range are excluded from the *SSD* efficient set implies that the probability of low returns is greater with these portfolios than with at least one *SSD* efficient portfolio with a higher mean and variance. Thus the application of the more restrictive *EV* rule can lead a highly risk-averse investor to make choices which

are inconsistent with the maximization of expected utility.

In summary, we feel that our results support the conclusion that except for the highly risk-averse investor the choice between the more familiar mean-variance model and the theoretically superior stochastic dominance model for selecting efficient portfolios is not critical. Where risk aversion is strong, however, second and third degree stochastic dominance rules are more consistent with the maximization of expected utility than is the mean-variance rule.

### REFERENCES

- K. J. Arrow, "Comment on the Portfolio Approach to the Demand for Money and Other Assets," *Rev. Econ. Statist.*, Feb. 1963, Suppl., 45, 24-27.
- W. J. Baumol, "An Expected Gain—Confidence Limit Criterion for Portfolio Selection," *J. Manage. Sci.*, Oct. 1963, 10, 174-82.
- K. Borch, "A Note on Uncertainty and Indifference Curves," *Rev. Econ. Stud.*, Jan. 1969, 36, 1-4.
- W. Breen, "Specific Versus General Models of Portfolio Selection," *Oxford Econ. Pap.*, Nov. 1968, 20, 361-68.
- and J. Savage, "Portfolio Distributions and Tests of Security Selection Models," *J. Finance*, Dec. 1968, 23, 805-19.
- Y. Chou, *Statistical Analysis*, New York 1969.
- J. L. Evans and S. H. Archer, "Diversification and the Reduction of Dispersion: An Empirical Analysis," *J. Finance*, Dec. 1968, 23, 761-67.
- E. F. Fama, "The Behavior of Stock Market Prices," *J. Bus. Univ. Chicago*, Jan. 1965, 38, 34-105.
- M. S. Feldstein, "Mean-Variance Analysis in the Theory of Liquidity Preference and Portfolio Selection," *Rev. Econ. Stud.*, Jan. 1969, 36, 5-12.
- J. E. Gaumnitz, "Maximal Gains from Diversification," working paper #10, School of Business, University of Kansas, June 1968.
- J. Hadar and W. R. Russell, "Rules for Ordering Uncertain Prospects," *Amer. Econ. Rev.*, Mar. 1969, 49, 25-34.

- G. Hanoch and H. Levy, "The Efficiency Analysis of Choices Involving Risk," *Rev. Econ. Stud.*, July 1969, 36, 335-46.
- M. C. Jensen, "Risk, The Pricing of Capital Assets and the Evaluation of Investment Portfolios," *J. Bus. Univ. Chicago*, Apr. 1969, 42, 167-247.
- H. Levy and G. Hanoch, (1970a) "Relative Effectiveness of Efficiency Criteria for Portfolio Selection," *J. Finance Quant. Anal.*, Mar. 1970, 5, 63-76.
- H. Levy and M. Sarnat, (1970b) "Alternative Efficiency Criteria: An Empirical Analysis," *J. Finance*, Dec. 1970, 25, 1153-58.
- J. Lintner, (1965a) "The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets," *Rev. Econ. Statist.*, Feb. 1965, 47, 13-37.
- , (1956b) "Security Prices, Risk, and Maximal Gains from Diversification," *J. Finance*, Dec. 1965, 20, 587-615.
- H. M. Markowitz, *Portfolio Selection*, New York 1959.
- , "Portfolio Selection," *J. Finance*, Mar. 1952, 12, 77-91.
- J. Mossin, "Optimal Multiperiod Portfolio Policies," *J. Bus. Univ. Chicago*, Apr. 1968, 41, 215-29.
- R. B. Porter, "Application of Stochastic Dominance Principles to the Problem of Asset Selection Under Risk," unpublished doctoral dissertation, Purdue Univ., Jan. 1971.
- J. W. Pratt, "Risk Aversion in the Small and in the Large," *Econometrica*, Jan.-Apr. 1964, 32, 122-36.
- J. P. Quirk and R. Saposnik, "Admissibility and Measurable Utility Functions," *Rev. Econ. Stud.*, Feb. 1962, 29, 140-46.
- M. K. Richter, "Cardinal Utility, Portfolio Selection and Taxation," *Rev. Econ. Stud.*, June 1960, 27, 152-66.
- W. F. Sharpe, "Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk," *J. Finance*, Sept. 1964, 19, 425-42.
- , "A Simplified Model for Portfolio Analysis," *J. Manage. Sci.*, Jan. 1963, 9, 277-93.
- J. Tobin, "Liquidity Preference as Behavior Towards Risk," *Rev. Econ. Stud.*, Feb. 1958, 26, 65-86.
- , "Comment on Borch and Feldstein," *Rev. Econ. Stud.*, Jan. 1969, 36, 13-14.
- G. A. Whitmore, "Third-Degree Stochastic Dominance," *Amer. Econ. Rev.*, June 1970, 60, 457-59.
- "The Fortune Directory: The 500 Largest U.S. Industrial Corporations," *Fortune*, July 1965, 72, 149-68.

# Progression and Leisure

By M. G. ALLINGHAM\*

It is well known that a tax on income may have either an incentive or a disincentive effect on the work of an individual, but there remains some confusion as to whether an increase in the degree of progression of such a tax which leaves tax revenue unchanged has an incentive or disincentive effect. The conventional theory was that, since such a change would raise the marginal tax rate without altering the average rate, there would typically be a decrease in the substitution effect with no change in the income effect, so the total effect would be negative and work would fall. This argument is incorrect because it confuses *ex ante* and *ex post* interpretations of equal yield, as has been shown for a particular case by Robin Barlow and Gordon Sparks, and by J. G. Head.

However, not only is the analysis of these authors overly restrictive, but it is also misleading in that it develops a condition for progression to be work-detering which depends on the measurement of utility, and it uses an unsatisfactory indicator of the degree of progression. This is shown in the present paper, which also develops a more satisfactory condition in the general framework. I commence with a statement of the model, discuss the existing analysis then present a more general analysis, and finally offer some conclusions.

## I. The Model

Consider a simple one-period model for an individual, who consumes an amount  $x$  of a perfectly divisible composite consumption commodity in the period (we may aggregate commodities as prices are fixed), and works a proportion  $y$  of the period. By appropriate choice of units we set his wage as one consumption unit per period, so that as he derives all his income from labor his pre-tax income in the period is  $y$  consumption units; clearly  $x$  and  $y$  are in the unit interval

$[0, 1]$ . The individual has a preference pre-ordering over combinations of  $x$  and  $y$  which may be represented by a real valued utility function  $u(x, y)$ . The utility function  $u$  has continuous second derivatives, is strictly increasing in  $x$  and  $1-y$ , strictly quasiconcave, and unbounded from below as  $x$  and  $1-y$  tend to zero; these are standard assumptions, the last reasonably ruling out no consumption and no leisure at the optimum. The proportion of positive income  $y$  payable in tax is  $t(y)$  (with some fixed payment for zero income). The rate of tax,  $t$ , has continuous second derivative, with the properties that tax payment is positive but less than income, and that the amount of income retained at the margin does not increase with income, that is

$$\partial^2(y(1-t))/\partial y^2 = -2t' - yt'' \leq 0$$

These are the weakest conditions on  $t$  which guarantee a nonempty compact choice set and thus the existence of an optimum. A tax is said to be progressive if its rate  $t$  increases with income and regressive if this decreases, so we define the *degree of progression*  $\pi$  for an income  $y$  to be  $t'(y)$ . Finally, the individual chooses his work level  $y$  to maximize his utility function  $u$  subject to his consumption not exceeding his post-tax income, that is subject to

$$x \leq y(1-t)$$

## II. Existing Analysis

The existing analysis of Barlow and Sparks imposes two severe restrictions: first, that  $u$  is additively separable in  $x$  and  $y$ ; and second, that the tax is levied at a constant positive rate  $\tau$  on incomes above some exemption level  $\epsilon$ , so for incomes above this level, the tax payment (or revenue)  $T$  is  $\tau(y-\epsilon)$ , and the tax rate is

$$t(y) = \tau - \frac{\epsilon\tau}{y}$$

\* University of Essex and Northwestern University.

Under these restrictions Barlow and Sparks derive an expression for the change in the optimal choice of work,  $\bar{y}$ , resulting from a small revenue-compensated change in the exemption levels  $\epsilon$ , that is  $d\bar{y}/d\epsilon|_{d\tau=0}$ , and show that this is positive if, and only if,

$$\frac{u_{yy}}{u_y} - \frac{u_{xx}}{u_x} < \frac{\tau}{\bar{y} - \epsilon},$$

where (here and throughout) all partial derivatives are evaluated at their equilibrium values. As might be expected, this condition clearly depends on the measurement of utility: if we apply an arbitrary increasing monotonic transform  $\phi(u)$  we add

$$\frac{\phi''}{\phi'} (u_y - u_x)$$

to the left of this inequality, so by appropriate choice of the arbitrary  $\phi''$  we may make the condition hold or not hold at will (note that both  $u_y$  and  $-u_x$  are negative so their sum is nonzero).

More important, there is in general no qualitative relationship between the effect on work of a change in the exemption level  $\epsilon$ , for which the above condition is developed, and the effect of a change in the degree of progression  $\pi$ , with which we are concerned. Here we have  $\pi = \epsilon\tau/\bar{y}^2$ , so an uncompensated increase in  $\epsilon$  is always associated with an increase in  $\pi$ , but a revenue-compensated change need not be. Such compensation requires that

$$dT = (\bar{y} - \epsilon)d\tau + \tau(d\bar{y} - d\epsilon) = 0,$$

or, using the expression for the total derivative of  $\bar{y}$ ,

$$d\tau = \frac{\tau(1 - \partial y/\partial \epsilon)}{y - \epsilon + \partial y/\partial \tau} d\epsilon;$$

then as

$$d\pi = \frac{\tau}{\bar{y}^2} d\epsilon + \frac{\epsilon}{\bar{y}^2} d\tau$$

the sign of  $d\pi/d\epsilon$  depends on the sign of

$$1 + \frac{\epsilon(1 - \partial y/\partial \epsilon)}{\bar{y} - \epsilon + \tau \partial y/\partial \tau}$$

Now  $\partial y/\partial \tau$  is the familiar effect on work of an increase in a proportional tax, and, even given  $\partial y/\partial \epsilon$ , is not bounded from above or below; thus if  $\epsilon$  is nonzero this expression, and therefore  $d\pi/d\epsilon$ , is not signed. Then as

$$\frac{d\bar{y}}{d\pi} = \frac{d\bar{y}}{d\epsilon} \frac{d\epsilon}{d\pi}$$

the sign of  $d\bar{y}/d\epsilon$  is in general irrelevant for the determination of the sign of  $d\bar{y}/d\pi$ , and thus for the progression problem.

Finally, we may note that the alternative to  $\pi$  as a definition of the degree of progression is  $-\frac{1}{2}\partial^2 x/\partial y^2$ , since a tax system may be considered as progressive if the extra consumption available from a small increase in income decreases with income, and regressive if this increases. It is clear that in Barlow and Sparks' model this is identically zero, so also irrelevant.

### III. General Analysis

In the general case we have noted that there exists a regular optimum, that is a solution where the budget constraint is satisfied as an equality and the amounts of consumption,  $x$ , work,  $y$ , and leisure,  $1-y$ , are strictly positive. We may then readily obtain the necessary conditions for an optimum as

$$\begin{aligned} u_x - \lambda &= 0 \\ u_y - \lambda(1 - t - \bar{y}t') &= 0 \\ x - y(1 - t) &= 0, \end{aligned}$$

where  $\lambda$  is a Lagrangean multiplier; this may be solved for the optimum work level  $\bar{y}$ . Using Taylor's expansion and omitting higher order terms we may write the tax rate in the neighborhood of this as

$$t(v) = t(\bar{y}) + (y - \bar{y})t'(\bar{y}),$$

or as

$$t = \alpha + \beta y,$$

where  $\alpha$  and  $\beta$  are appropriate constants (given  $\bar{y}$ , which for notational convenience we now write as  $y$ ); from the constraints on  $t$ ,  $\beta$  must be nonnegative. Since we are examining the effects of *small* changes in the degree of progression, considering only the neighborhood of the equilibrium involves no

loss of generality. We have, however, added the necessary dimension to the problem for tax revenue is now locally quadratic in income, whereas in Barlow and Sparks' case, it is locally linear. A manifestation of this is that the two possible definitions of the degree of progression, instead of being qualitatively inconsistent, now become identical, for locally  $t'$  is simply  $-\frac{1}{2}\partial^2 x/\partial y^2$ ; both are equal to  $\beta$ . We may then conveniently consider the tax rate as comprising two parts: a *flat rate*  $\alpha$ , and a *progressive part*  $\beta y$  (with degree of progression  $\beta = \pi$ ); equivalently the flat rate ( $\alpha$ ) is simply the difference between the average rate ( $t = \alpha + \beta y$ ), and the excess of the marginal over the average rates ( $\mu - t = \beta y$ ).

The sufficient condition for an optimum is that the bordered Hessian

$$\begin{bmatrix} 0 & 1 & \mu-1 \\ 1 & u_{xx} & u_{xy} \\ \mu-1 & u_{yx} & u_{yy}-2\beta\lambda \end{bmatrix}$$

be negative definite under constraint; in this  $\mu$  is  $\alpha + 2\beta y$ , which is simply the marginal rate of tax,  $\partial T/\partial y$ .

By differentiating the equilibrium conditions with respect to  $\alpha$  and solving we may obtain in the usual way the effect of a change in  $\alpha$  on work,

$$\frac{\partial y}{\partial \alpha} = -\frac{y\kappa}{\delta} - \frac{\lambda}{\delta}$$

where  $\kappa$  is the cofactor of the element  $\mu-1$  in, and  $\delta$  the determinant of, the above bordered Hessian. From the necessary and sufficient conditions, respectively,  $\lambda$  and  $\delta$  are positive, and we may interpret the two terms on the right of this equality as the traditional unsigned substitution effect and negative income effect respectively. We may similarly obtain the effect of a change in  $\beta$ ,

$$\frac{\partial y}{\partial \beta} = -\frac{y^2\kappa}{\delta} - \frac{2\lambda y}{\delta},$$

which we may interpret in the same way.

Now as the degree of progression  $\pi$  is equal to  $\beta$ , we may investigate the effect of a change in this by considering a revenue-compensated change in  $\beta$ . Proceeding as in

the restricted case we see that such compensation requires that

$$dT = yd\alpha + y^2d\beta + \mu dy = 0$$

or that

$$d\alpha = -\frac{y^2 + \mu\partial y/\partial \beta}{y + \mu\partial y/\partial \alpha} d\beta;$$

then using the expression for the total derivative of  $y$  and the expressions obtained above for  $\partial y/\partial \alpha$  and  $\partial y/\partial \beta$  we have, as  $\beta$  is equal to  $\pi$ ,

$$\frac{dy}{d\pi} \bigg|_{dT=0} = \frac{-\lambda y^2/\delta}{y + \mu\partial y/\partial \alpha}$$

This is the general expression for the effect on work of a revenue-compensated increase in the degree of progression. Since  $\lambda$  and  $\delta$  are positive the numerator of this expression is unambiguously negative, so the whole expression is negative if the denominator is positive, that is if, and only if,

$$\frac{\partial y}{\partial \alpha} \frac{\alpha}{y} > -\frac{\alpha}{\mu}$$

(assuming that  $\alpha$ , and thus  $\alpha/\mu$ , is positive; in the unlikely event that  $\alpha/\mu$  is negative this inequality is reversed and the following interpretation amended accordingly). The expression on the left of this inequality is simply the elasticity of work,  $y$ , with respect to the flat rate of tax,  $\alpha$ , and that on the right minus the ratio of the flat to the marginal rates of tax. This ratio lies between zero and unity, as the denominator is the positive numerator plus some nonnegative amount  $2\beta y$ ; further, it is equal to unity when there is no progression, and decreases as the degree of progression,  $\pi$  or  $\beta$ , increases. The above condition is purely in terms of observable quantities, and thus clearly independent of the measurement of utility.

#### IV. Conclusions

We may conclude from the above analysis that a small revenue-compensated change in the degree of tax progression has a disincentive effect on work if, and only if, the elasticity of work with respect to the flat rate of tax exceeds minus the ratio of the flat to the

marginal rates of tax. If the original tax is purely proportional the condition is simply that the elasticity of work with respect to this tax exceed minus unity; as the degree of progression increases, the effects of further increases are *less* "likely" to be work-detering. However increases in the degree of progression will always be work-detering if increases in proportional taxes are work-stimulating, and in a loose sense unless these are significantly work-detering.

We have then identified reasonably interpretable necessary and sufficient conditions for progression to be work-detering in the general case. These confirm that the conventional result does not hold in general, and that while it seems unlikely that it should not hold, the conditions for this are far less pathological than has been suggested, for example by Barlow and Sparks and by Head.

While this result is of interest in its own right, it may also be of value in the construc-

tion of optimal income tax schedules. It is, however, important to note that it does not have any *immediate* implications here: firstly because the optimal tax problem involves aggregate constraints and thus requires a general equilibrium approach, and secondly because a decrease in work may be socially beneficial—if the social evaluation of the resulting increase in leisure outweighs that of the decrease in consumption. Thus the result that progression is typically work-detering is not an argument against progressive taxation.

#### REFERENCES

- R. Barlow and G. R. Sparks, "A Note on Progression and Leisure," *Amer. Econ. Rev.*, June 1964, 54, 372-77.  
J. G. Head, "A Note on Progression and Leisure: Comment," *Amer. Econ. Rev.*, Mar. 1966, 64, 172-79.

# Peasants, Procreation, and Pensions: Comment

By MARIANNE ABELES FERBER\*

In a recent article in this *Review*, Philip Neher expresses concern with the problem of overpopulation and suggests that an important contributing factor may be the attempt to provide for old age by raising children. He concludes that "if a population control problem is thought to exist, a control scheme might well include the provision of pensions . . .".

Neher's reasoning seems plausible, as far as it goes. He overlooks, however, an additional approach that might be even more effective.

At present, a large proportion of married females is still unemployed, many of those employed have lower-level jobs than men

with comparable qualifications, and most are paid less when doing the same job as men. Because of this situation the opportunity cost of raising children, in purely economic terms, is relatively low. To this may be added the fact that the unemployed or underemployed woman may also turn to child rearing for psychological reasons.

Hence, one may conclude that providing equal, or at any rate, greater opportunity for women in the labor market and removing or reducing tabus on working wives is likely to be an effective way to increase the ratio of costs to benefits involved in child rearing.

## REFERENCES

- P. A. Neher, "Peasants, Procreation, and Pensions," *Amer. Econ. Rev.*, June 1971, 61, 380-89.

\* Assistant professor of economics, University of Illinois at Urbana.

# Peasants, Procreation and Pensions: Reply

By PHILIP A. NEHER\*

Marianne Ferber complains that I overlooked the economic status of women as a population control device. A brief review of my article should convince her that my neglect stems from having assumed that women in the model economy had already achieved a good measure of economic and social equality. In fact, the role of women cannot be distinguished from that of men. That is super-equality for you, which probably cannot be supported by biological, much less social or economic, fact. It was simply a convenient assumption.

Ferber concludes that improved status for women would reduce the desired size of family by increasing the opportunity cost of bearing and raising children. By emphasizing this *substitution effect*, she neglects the *income effect*. If children are a normal (superior)

good, the income effect may swamp the substitution effect. Greater opportunity for women make children more "expensive," but more can be "afforded." Which effect wins out is, of course, an empirical matter.

Ferber is on firm ground if she wants to conceive (of) children as inferior goods, for then, the income and substitution effects work in the same direction to reduce the desired family size.<sup>1</sup>

## REFERENCES

- M. J. Brennan, *Theory of Economic Statics*, Englewood Cliffs 1970.  
P. A. Neher, "Peasants, Procreation, and Pensions," *Amer. Econ. Rev.*, June 1971, 61, 380-89.

<sup>1</sup> My analysis parallels Michael Brennan, p. 317. Substitute "rearing children" for "leisure" and assume fixed coefficients linking rearing time and numbers of children.

\* University of British Columbia

# Upward Sloping Demand Curves Without the Giffen Paradox

By DANIEL C. VANDERMEULEN\*

It has long been known that the axioms of consumer theory do not rule out upward sloping demand curves so long as the second-order conditions are expressed solely by a sign convention on the bordered Hessian determinant. Only bordered principal minors have determinate signs, so any expression involving off-diagonal minors must have an ambiguous sign. But economists have been uneasy with this indeterminate result and have tried, short of strengthening the axioms, to narrow the range of uncertainty or, at least, to minimize its empirical significance.

The most notable attempt is that of J. R. Hicks, who attributes a positive price effect to a *high percentage of income* spent on an inferior commodity and, in the following well-known passage, appraises the likelihood of its occurrence:

Consumers are only likely to spend a large proportion of their incomes on what is for them an inferior good if their standard of living is very low. The famous Giffen case . . . exactly fits these requirements. . . . But it is evident how rare such cases must be. . . . Thus, as we might expect, the simple law of demand—the downward slope of the demand curve—turns out to be almost infallible in its working. [p. 35]

Later writers have accepted Hicks' conclusion and, particularly in intermediate textbooks, have worded it even more strongly.<sup>1</sup> It has become part of economic dogma that upward sloping demand curves are rare because Giffen cases are rare.

In Hicks' argument there are dubious points that have apparently gone unde-

tected. In the Slutsky equation the income effect of a price change depends on the number of units purchased, not the percentage of income spent. It is in Alfred Marshall's treatment that a high percentage of income spent on a commodity is crucial, since a change in price then affects the marginal utility of money. Marshall argued that a rise in price of some staple like bread might so raise the marginal utility of money as to reduce the quantities of competing food-stuffs and increase the consumption of bread.<sup>2</sup> But in the Slutsky equation the marginal utility of income appears in the pure substitution effect. Thus, there is no assurance that the Giffen conditions, on balance, yield a dominant positive income term.

Actually, a much stronger assertion can be made, once the problem is correctly reformulated. The Giffen case must be purged of what is, unfortunately, its most persuasive feature, any implication that devoting a high percentage of an abnormally low real income to a single commodity influences a consumer's preferences. With money income and the price of the composite commodity constant, both the amount spent on a commodity and a positive price effect depend solely on the properties of the indifference map. Thus, the Hicksian hypothesis is that exceptional regions of consumer preference are likely to occur only at lower levels of utility and to be revealed only by abnormally low levels of real income. So stated, the argument simply does not hold up. Section I tests the Hicksian hypothesis by constructing a specific utility function to yield a positive price effect. The function shows that the Giffen case, far from being necessary, actually represents the set of conditions least likely to generate an upward sloping demand curve. Section II

\* Professor of economics, Claremont Graduate School.

<sup>1</sup> For example, C. E. Ferguson states: "The law of demand fails to hold only in the case of Giffen's paradox . . ." (p. 57).

<sup>2</sup> For further analysis, see George Stigler and William Gramm.

generalizes from the specific utility function to the basic pattern of consumer preference that generates positive price effects.

### I. A Specific Function

With  $x_2$  a composite commodity and  $p_1$  declining, the geometric requirement for a backward sloping price-consumption curve is that indifference curves flatten out along vertical lines more rapidly than the rotating budget line. Expressed analytically the geometric requirement is:<sup>3</sup>

$$(1) \quad \frac{\partial}{\partial x_2} \left( -\frac{p_1}{p_2} \right) = \frac{1}{x_1} < \frac{(f_1 f_{22} - f_2 f_{12})}{f_2^2} \\ = \frac{\partial}{\partial x_2} \left( -\frac{f_1}{f_2} \right)$$

Since  $f_1 = \lambda p_1$  and  $f_2 = \lambda p_2$ , the inequality is the familiar condition that the income effect of a price change be positive and exceed the pure substitution effect in absolute value. The geometric requirement is, then, both necessary and sufficient for a positive price response.

To fulfill the geometric requirement, tangent lines to higher indifference curves must have lower vertical intercepts when  $x_1$  is constant; but to satisfy convexity, tangents to the same indifference curve must intersect the vertical axis in higher points as  $x_1$  decreases. Both conditions are met when the tangent line through any point  $(x_1, x_2)$  has the vertical intercept:

$$(2) \quad b(x_1, x_2) = kx_1^m x_2^{-n} + nx_2(n+1)^{-1},$$

where  $k > 0$ ,  $m < 0$ ,  $-1 < n$ . The slope of the indifference contour through any point satisfies the following equation:

$$(3) \quad \frac{dx_2}{dx_1} = \frac{x_2 - b(x_1, x_2)}{x_1} \\ = \frac{(n+1)^{-1} x_2^{n+1} x_1^{-n-2} - kx_1^{m-2}}{x_2^n x_1^{-1}}$$

The preceding exact differential equation has the solution:

<sup>3</sup> From the budget equation,  $-p_1/p_2 = x_2/x_1 - y/p_2 x_1$ .

$$U = f(x_1, x_2)$$

$$(4) \quad = (n+1)^{-1} x_2^{n+1} x_1^{-1} - (1-m)^{-1} k x_1^{m-1}$$

The partial derivatives are:

$$f_1 = x_1^{-2} [-(n+1)^{-1} x_2^{n+1} + kx_1^m];$$

$$f_2 = x_1^{-1} x_2^n; \quad f_{12} = -x_2^{n-2};$$

$$(5) \quad f_{11} = x_1^{-3} [(n+1)^{-1} 2x_2^{n+1} + k(m-2)x_1^m];$$

$$f_{22} = nx_1^{-1} x_2^{n-1}$$

The function (4) and all first and second partial derivatives exist and are continuous for  $x_1 > 0$ ,  $x_2 \geq 0$ . The marginal utility of  $x_2$  is nonnegative everywhere on these ranges, but  $f_1 \geq 0$  holds only for:

$$(6) \quad x_2^{n+1} x_1^{-m} \leq (n+1)k$$

The equality in (6) is the boundary of satiation for  $x_1$ , the locus of points at which the indifference curves begin to slope upward. The arbitrary constant  $k$  shifts the threshold of satiation to any desired level. Below the boundary of satiation, (4) is clearly an increasing function. The region where  $f_1 < 0$  does not affect the analysis but can be eliminated, if desired, by a free disposal assumption that extends the indifference curves horizontally.

Even below the threshold of satiation, (4) is not everywhere quasi-concave. The required nonnegativity of the bordered Hessian determinant traces out a lower boundary for convexity.<sup>4</sup> Indifference curves that pass through the boundary have points of inflection on it, bend downward, and intersect the horizontal axis. For the present, all indifference contours with concave arcs can be ig-

<sup>4</sup> After expansion, the requirement for a nonnegative bordered Hessian determinant becomes:  $y^2 - [2 - (n+1)n^{-1}m]y + 1 \leq 0$ , where  $y = x_2^{n+1} [(n+1)kx_1^m]^{-1}$ . The conditions holds for  $r_1 \leq y \leq r_2$ , where the roots are real, unequal, positive, and with  $r_1 r_2 = 1$ . Transformed back the requirement for convexity is  $r_1(n+1)k \leq x_2^{n+1} x_1^{-m} \leq r_2(n+1)k$ . Since  $r_2 > 1$ , the upper limit lies beyond the boundary of satiation and is inoperative. The lowest indifference curve to lie entirely within the region of convexity is  $x_2^{n+1} x_1^{-m} = (n+1)k(1-m)^{-1}$ , when  $U = 0$ . Written as  $y = 1/(1-m)$ , it everywhere satisfies the second-order conditions. For a convenient numerical example, set  $n = 3$ ,  $m = -1$ .

nored and the boundary fixed at the lowest indifference curve that is everywhere convex. The bordered Hessian determinant is always positive, and (4) satisfies all other requirements for a utility function, on the region:

$$(7) \quad (n+1)k/(1-m) \leq x_2^{n+1} x_1^{-m} \leq (n+1)k$$

With  $|H|$  denoting the bordered Hessian determinant and for  $n > 0$ , the pure income and the price effects on  $x_1$ , and the cross effect on  $x_2$ , are, respectively:

$$(8) \quad \frac{\partial x_1}{\partial y} = \frac{p_2 f_{12} - p_1 f_{22}}{|H|} = - \frac{x_1^{-2} x_2^{n-1} (p_2 x_2 + n p_1 x_1)}{|H|} < 0$$

$$(9) \quad \frac{\partial x_1}{\partial p_1} = \frac{-\lambda p_2^2 + x_1(-p_2 f_{12} + p_1 f_{22})}{|H|} = \frac{n p_1 x_2^{n-1}}{|H|} > 0$$

$$(10) \quad \frac{\partial x_2}{\partial p_1} = \frac{\lambda p_1 p_2 + x_1(p_2 f_{11} - p_1 f_{12})}{|H|} = \frac{m k p_2 x_1^{m-2}}{|H|} < 0$$

The slope of the price-consumption curve for  $x_1$  is:

$$(11) \quad \frac{dx_2}{dx_1} = \frac{\partial x_2}{\partial p_1} / \frac{\partial x_1}{\partial p_1} = \frac{k m p_2 x_1^{m-2}}{n p_1 x_2^{n-1}} < 0$$

So long as  $n > 0$  and  $m < 0$ , the utility function (4) generates a negative sloping arc in the price-consumption curve that bends upward, has no inflection point, and approaches a vertical slope as  $p_1 \rightarrow 0$ . The corresponding arc of the demand curve has a positive slope, bends downward, and intersects the horizontal axis in a vertical slope.<sup>5</sup> See Figure 1.

<sup>5</sup> In his article, H. H. Liebhafsky presented a specific utility function showing a negative pure income effect, the first such function, he believed, to appear in print. He also challenged readers to carry the analysis further by devising a similar function with nonvanishing cross

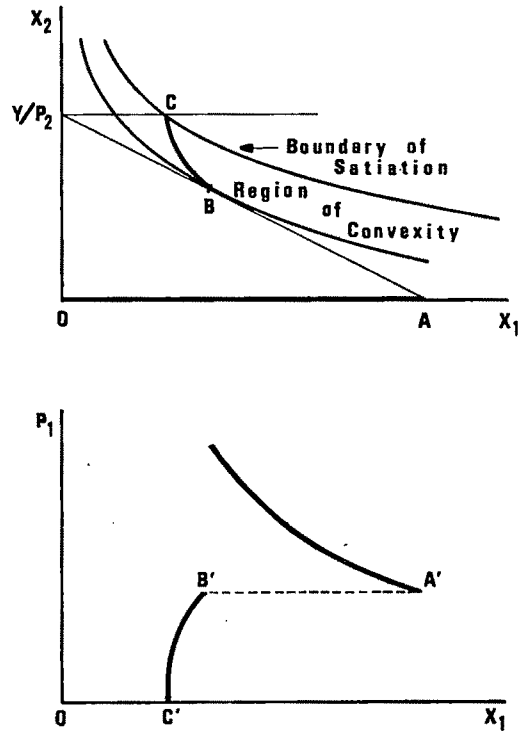


FIGURE 1. PRICE-CONSUMPTION AND DEMAND CURVES

It has become firmly entrenched in the economic literature that a high percentage of income spent on a commodity is a necessary, or at least a predisposing, condition for an upward sloping demand curve. The preceding results show that the share of income is simply irrelevant. With the appropriate choice of income and prices, the initial equilibrium can occur, and the backward sloping price-consumption arc can begin, at any point on the indifference curve that forms the lower boundary in (7). Hence, the percentage of income spent on  $x_1$  may be chosen at will.

The second Giffen condition, that a consumer have an abnormally low real income, fares even worse; it tends to prevent upward sloping demand curves. Since a consumer must be buying a positive amount of a com-

derivatives. At the August 1968 meetings of the Western Economic Association, I presented a version of the utility function (4). It meets Liebhafsky's challenge and generates an upward sloping demand curve as well.

modity before buying less, the preference pattern or utility function that generates a positive price effect cannot be extended to the origin. Convexity breaks down. Below the region defined by (7), all indifference curves cross a lower boundary, become concave, and intersect the horizontal axis. A sufficiently low budget line touches only concave arcs and has a corner optimum on the axis. As  $p_1$  declines, a consumer continues to spend all income on  $x_1$  until the budget line reaches tangency with a convex arc. The price-consumption locus  $OABC$  in Figure 1 coincides with the horizontal axis until it jumps discontinuously up and back to an upward bending arc. The demand curve has unit elasticity down to  $A'$ , when it shifts discontinuously leftward to a concave arc  $B'C'$ . Perhaps this violently discontinuous reaction is a more realistic prediction of behavior under extreme poverty, but it is at odds with the traditional smooth version.

A positive price effect is always possible at a sufficiently high level of real income but there must exist a region of the indifference map near the origin in which such an effect cannot occur. By choice of an appropriately small value for  $k$  in the utility function (4), the region near the origin can be shrunk to any desired size, but only at the cost of a drastic lowering in the threshold of satiation. Perhaps the intuitive appeal of the historical Giffen case is that it makes a low threshold of satiation plausible. But, then, its significance lies in showing that an upward sloping demand curve can occur *even* at abnormally low levels of real income.

## II. A General Pattern

At the very least, the utility function (4) is a counter example that casts doubt on the linkage of a positive price effect to the Giffen case. It has, however, been derived directly from the necessary and sufficient geometric condition for a positive price effect, and any other specific function so derived is likely to support the conclusions. To complete the indifference map, (4) has to be spliced with another function at lower levels of utility or embedded in a more general pattern of preference that satisfies

the accepted axioms everywhere on the consumption set. This general pattern supplies the basic requirements for upward sloping demand curves.

As just demonstrated, the price effect cannot be positive everywhere on the consumption set but must take on negative values as well.<sup>6</sup> But  $\partial x_1/\partial p_1$  is derived by differentiation of the first-order conditions and inherits from the second partials of the utility function the property of continuity. Hence,  $\partial x_1/\partial p_1$  must pass continuously from negative to positive values. Geometrically, the *RCS* must decline along vertical lines, first more slowly, then more rapidly than the price ratio of the upward rotating budget line. Thus, the general pattern, of which a positive price effect is the culmination, is an increasingly rapid rate of decline in the *RCS* along vertical lines, or more generally along rays.

A commodity will be termed *relatively inferior* when its *RCS* with respect to the composite commodity declines along a ray. If the decline in the *RCS* is monotonic along all rays, the property of relative inferiority holds over the entire indifference map and becomes an attribute of the commodity. When the *RCS* declines at a sufficiently rapid rate along rays it also declines along vertical lines. The commodity becomes inferior in the usual sense, but only within a region of the consumption set; at sufficiently low levels of income every commodity must be normal. A still more rapid decline in the *RCS* along rays yields a positive price effect.

A relatively inferior commodity has an income elasticity less than one, so increases in income reduce the percentage spent on it. If the *RCS* declines along rays at a sufficiently rapid rate, a consumer initially responds to a growing income by increasingly larger reductions in the percentage spent on the commodity, and then continues the pattern by increasingly larger reductions in the quantity purchased. A further growth of income ultimately opens up a region of the

<sup>6</sup> The price effect cannot be zero everywhere since vertical price-consumption curves make a right-angle turn at the horizontal axis.

indifference map where the consumer reduces purchases even when the price falls. Thus, a positive price effect, rather than requiring a high share of income, is likely to occur only after a consumer has been reducing at an increasingly rapid rate the percentage of income spent on a commodity. In the context of the whole indifference map this Giffen condition, like the other, turns out to be not merely irrelevant but counter-productive.

The price-consumption curve provides further insight. So long as indifference curves intersect the vertical axis, price-consumption curves must slope downward at the axis.<sup>7</sup> With the price effect negative, as it must be initially, (11) shows that the slope depends on the sign of the cross effect. When  $x_1$  is small, the income effect on  $x_2$  of a change in  $p_1$  is also small, and  $x_2$  is a gross substitute. The increasing relative inferiority of  $x_1$ , together with the larger quantity purchased, converts  $x_2$  into a gross complement and causes the price-consumption curve for  $x_1$  to slope upward. When the price effect goes positive, the curve bends backward and terminates in a point of satiation. A pattern of steadily increasing relative inferiority causes the price-consumption curve to end in a concave arc rather than the convex arc generated by (4). The demand curve curls backward toward the vertical axis.

The ingredients for a positive price effect are inferiority and a sufficiently rapid approach toward satiation—the rate of decline in the *RCS* along vertical lines must be such that, if continued, satiation would occur.<sup>8</sup> Inferiority alone, however great, does not imply satiability. For a geometric verification it is only necessary to construct the upper portion of an indifference map by

<sup>7</sup> The economic implications of non-intersection are unacceptable; deprived of  $x_1$  a consumer is indifferent with respect to the quantities of all remaining  $(n-1)$  commodities.

<sup>8</sup> If the decline in the *RCS* is not monotonic, presumably drastic reductions could occur locally many times without terminating in satiation. It is not clear how much variability in the *RCS* the axioms and the continuity requirements permit. Repeated reversals of the pattern of preference do, however, become less plausible.

drawing an indifference curve asymptotic to the horizontal axis and shifting it upward parallel to itself along backward sloping parallel lines. With prices constant the consumer ceases to buy  $x_1$  at a moderate level of income but will then expand purchases without limit as  $p_1$  falls with income constant. At the other extreme a consumer may reach satiation at a small quantity of  $x_1$  without its ever becoming inferior. A specific example occurs when the *RCS* declines along each ray exactly as fast as the price ratio of some rotating budget line. Each ray becomes a price-consumption locus up to the point of satiation at which the budget line goes horizontal and the *RCS* reaches zero.<sup>9</sup>

Inferiority directly increases the income effect of a price change, but the pure substitution effect is inversely related to the curvature of the indifference contours. By increasing the curvature, a rapid approach toward satiation depresses the pure substitution effect.<sup>10</sup> From the geometric requirement expressed in (1), both inferiority and satiability are necessary for a positive price effect, but to be sufficient they must be present in the proper combination. For example, when  $n \leq 0$ , the utility function (4) may generate both a negative income effect and a boundary of satiation, but the demand curve for  $x_1$  does not slope upward.

The traditional economic explanation of upward sloping demand curves has erred in focusing on the income term alone and in not identifying the conditions that influence the relative size of both the income and substitu-

<sup>9</sup> The utility function,  $U = \log (x_2/x_1) - kx_1^{-1}$ , where  $k > 0$ , satisfies the geometric conditions. The numerator of  $\partial x_1/\partial y$  is  $p_1 x_2^{-2} > 0$ .

<sup>10</sup> As the optimum moves along an indifference curve with  $p_2$  constant,

$$\frac{d}{dx_1} \left( \frac{p_1}{p_2} \right) = \frac{1}{p_2} \frac{dp_1}{dx_1} = \frac{d}{dx_1} \left( \frac{f_1}{f_2} \right) = - \frac{|H|}{p_2^2 f_2},$$

or

$$\frac{dx_1}{dp_1} = - \frac{p_2 f_2}{|H|} = - \frac{\lambda p_2^2}{|H|}$$

A rapid approach to satiation implies a large absolute value for  $d(f_1/f_2)/dx_1$  and a small pure substitution effect.

tion terms. Once this is done, it is clear that a positive price effect is more likely to be revealed by growing affluence than by extreme poverty. Thus, the virtual identification of a positive price effect with the Giffen case has served economists badly. It has directed attention toward the least likely of empirical conditions and has given false assurance of the near infallibility of the law of demand.

#### REFERENCES

- C. E. Ferguson, *Microeconomic Theory*, 2d ed., Homewood 1969.
- W. P. Gramm, "Giffen's Paradox and the Marshallian Demand Curve," *Manchester Sch. Econ. Soc. Stud.*, Mar. 1970, 38, 65-71.
- J. R. Hicks, *Value and Capital*, 2d ed., Oxford 1946.
- H. H. Liebhafsky, "New Thoughts About Inferior Goods," *Amer. Econ. Rev.*, Dec. 1969, 59, 931-34.
- A. Marshall, *Principles of Economics*, 8th ed., London 1920.
- G. J. Stigler, "Notes on the History of the Giffen Paradox," *J. Polit. Econ.*, Apr. 1947, 55, 152-56.

# Determinants of the Commodity Structure of U.S. Trade: Comment

By LAWRENCE WEISER AND KEITH JAY\*

Theoretical and empirical analyses of international commodity trade patterns have developed in several directions since Wassily Leontief's article on the validity of the Heckscher-Ohlin theory. Robert Baldwin's recent article in this *Review* provides a comprehensive survey of the important issues in contemporary trade literature. His efforts at location and consolidation of data from many sources have shed new light on several competing hypotheses attempting to explain trade patterns.

This paper will attempt to improve Baldwin's statistical results by combining his data with some developed by Gary Hufbauer in a parallel study of the industrial characteristics of trade. Also, the relationship between export performance and technological innovation will be explored by a new approach which focuses on actual change in the production function and not merely potential improvement by Research & Development (R&D) efforts. Finally, economic and industry characteristics of several countries other than the United States will be used to provide additional information concerning the relationship of technological change and economies of scale to trade flows.

## I

Baldwin presents the results of his regression analysis relating various economic characteristics of U.S. industries to the trade balances of those industries. The regression coefficients for the independent variables representing the proportions of R&D workers and some other skilled and unskilled categories in the industries' labor forces are statistically significant and have

the theoretically correct sign in many cases, but there are several exceptions as Baldwin admits (pp. 136-39). Especially disappointing from the standpoint of theoretical prediction and previous empirical studies is the failure of the economies of scale variable to show a significant positive relationship with the U.S. industries' trade balances. Furthermore, although one cannot expect very high  $R^2$  values for cross-section data of this type, perhaps they can be raised from the .20-.50 range reported by Baldwin.

The following regression equation will be used to represent a reasonable theoretical explanation of U.S. trade patterns and will yield good statistical results.

$$(1) \quad ES = a_0 + a_1R\&D + a_2 \text{Scale} + a_3S4 + a_4S5 + a_5S6$$

ES (export share) is the ratio of U.S. exports of each industry to the exports of the eleven leading exporting countries.<sup>1</sup> This measure replaces Baldwin's dependent variable, the trade balance of each industry expressed as a proportion of total U.S. trade. Variants of the export share have been

<sup>1</sup> Export data were taken from the *United Nations Statistical Papers, Series D*. The countries included in the calculations were the United States, United Kingdom, France, West Germany, Italy, Belgium-Luxembourg, Netherlands, Canada, Japan, Australia, and Sweden. These nations account for 65 percent of total world trade excluding the socialist bloc, and each country contributed at least 2 percent of world exports while no nation outside this group reached the 2 percent value (calculated from data reported in *United Nations Yearbook of International Trade Statistics*). The years used were 1960 and 1967. Manufactured exports were divided into thirty-two industrial categories with food processing, ordinance, and miscellaneous manufacturing excluded. The industrial classification follows very closely the sectors of the U.S. Department of Commerce 1958 input-output study, but is also consistent with SIC and SITC systems, so that other data reported on those categories could be utilized. The concordance between the three classification systems used will be provided by the authors on request.

\* Assistant professor of economics, University of Illinois, Urbana, and international economist, Agency for International Development, Washington, D.C., respectively.

used in previous empirical studies by William Gruber et al. and Donald Keesing as an index of an industry's ability to compete in world markets, although neither study employed multiple regression analysis. There are criticisms which can be made of this measure as an indicator of comparative advantage, and some of these are discussed by Gruber et al., pp. 22-25. Neglecting the import side of the trade balance would ordinarily bias the results due to inter-industry differences in tariffs and transportation costs. Also, substantial differences in demand structure between the United States and its trading partners might be reflected in the export/sales ratio of industries without any underlying comparative cost differences.

Normalizing *U.S.* exports by total exports for each industry satisfies these objections to a large extent because all these countries face similar interindustry differences in trade barriers, although *EEC* members have substantially lower absolute tariff levels for intra-union trade. The distortion due to international divergences in demand structure is moderated by focusing on the competitive ability of each *U.S.* industry vis-à-vis other advanced economies. The export share, although not a measure of traditional comparative advantage, was chosen on both theoretical and empirical grounds to be the best indicator of export performance. It provides a method of ranking industries which does not depend on their absolute or relative importance in the domestic economy as would using simply the value of exports or exports minus imports, or on their importance in total world trade as would exports divided by domestic output.<sup>2</sup>

All the independent variables except scale were taken directly from Baldwin's data and converted to the thirty-two industry classification system by weighting each industry by its output in the 1958

input-output table. *R&D* is the proportion of engineers and scientists in the labor force of each industry. Similar measures for low skill workers are *S4* operatives, *S5* nonfarm laborers and service workers, and *S6* farm laborers.

Scale is a variable which measures the economies of scale for each industry.<sup>3</sup> Industries which are capable of achieving high increases in value-added per worker as the size of the firms increase, should give countries with large domestic markets like the United States a competitive export advantage over smaller countries in those industries. Therefore *U.S.* industries with high values for scale should have large export shares, and  $\alpha_2$  should be positive. This scale variable is a definite improvement on Baldwin's measure which is based on the percentage of employees in establishments with 250 or more employees in each industry. His variable does not relate size to productivity differences as Hufbauer's clearly does.<sup>4</sup>

Table 1 presents the results of the regression analysis using *U.S.* trade performance in 1960, the year for which the industrial characteristics are calculated, and 1967, the most recent year for which trade data were available.

<sup>3</sup> The economies of scale variable was formulated and estimated by Hufbauer, pp. 178, 212-23. For each industry the scale elasticity  $\alpha$  is derived from the following equation:  $V = K(E/N)^\alpha$ . The coefficient  $V$  is the ratio between value-added per employee in a particular size plant and the average value-added per worker for all establishments in that industry;  $K$  is a constant;  $(E/N)$  is the average number of employees per establishment; and  $\alpha$  is the scale elasticity. Taking logarithms of both sides and regressing  $V$  on  $(E/N)$  for each size class of establishment in each industry gives estimates for  $\alpha$  for each industry. These estimates include negative values for diseconomies of scale and positive values for economies of scale. Hufbauer's results were converted to the thirty-two sector classification system by weighting each industry by its exports.

<sup>4</sup> The estimates of  $\alpha$  may have an upward bias because of systematic relationships between plant size and 1) product type, 2) quantity and quality of human and physical capital, and 3) monopoly power (see Hufbauer, pp. 179-81). However, when compared to estimates based on engineering data, values of  $\alpha$  are somewhat low. Despite these caveats, Hufbauer's scale measure is useful because of its broad coverage of industries, wide range of plant sizes, and clear indication of the relationship between size and productivity.

<sup>2</sup> Exports, exports minus imports, and exports divided by imports were regressed on the independent variables in equation (1), but the  $R^2$ s were substantially lower, and no statistically significant coefficients of the correct theoretical sign were obtained.

TABLE 1—REGRESSION RESULTS OF *U.S. EXPORT SHARES ON U.S. INDUSTRY CHARACTERISTICS, 1960, 1967*

Year	$a_0$	$a_1$	$a_2$	$a_3$	$a_4$	$a_5$	$R^2$	$F$
1960	30.1** (3.01)	3.11** (3.09)	8.07** (3.53)	-.372* (-2.13)	-.668* (-2.01)	.458* (2.16)	.78	18.15
1967	27.9** (3.07)	2.24* (2.46)	6.47** (3.12)	-.343* (-2.16)	-.537* (-1.78)	.289 (1.50)	.72	13.51

Note: *t*-values are given in parentheses. Number of observations equals 32.

\* Significant at .10 level.

\*\* Significant at .01 level

In almost all instances the regression coefficients have the theoretically expected sign and are statistically significant. The major exception is  $a_5$ , the coefficient for farm laborers. These are certainly low-skill workers, and therefore the United States should not have an advantage in exporting commodities which require a relatively large number of this group; these workers have the benefit of using high quantities of *U.S.* agricultural resources in conjunction with their labor. It is this combination which raises the productivity of these workers, and so the positive sign for  $a_5$  reflects the *U.S.* advantage in agricultural production rather than an abundance of farm laborers.

It should be noted that only high skill workers and low skill workers were represented as independent variables in the regression equation. Intermediate skill categories such as clerical, sales, craftsman, etc., are not likely to be crucial determinants of the *U.S.* trade pattern. Baldwin's calculations show similar requirements of these skill classes in *U.S.* exports and import replacements (p. 136, Table 2). However, omission of these variables could bias the statistical results due to correlation with one or more of the other independent variables. In order to guard against this possibility a step-wise regression program was used which automatically excluded any independent variable not meeting a .20 level significance test for its coefficient. The intermediate skill classes failed to meet this restriction and were omitted. As a final check on bias from this source, the additional skill categories

were forced into a regression equation similar to (1). The coefficients of the high and low skill variables and scale remained significant with the expected sign.

Compared to Baldwin's regression results, the explanatory power of the five variables in (1) as indicated by the values of  $R^2$  and  $F$  is quite high. Contributing to this improvement is the use of export share as the dependent variable and the substitution of Hufbauer's measure of economies of scale as one of the independent variables. The regression coefficient for scale  $a_2$  is positive and statistically significant at the .01 level. The success of this fairly concise equation in relating *U.S.* trade patterns to three major alternatives to the Heckscher-Ohlin theory<sup>5</sup> is quite encouraging with respect to formulating hypotheses which can be used to analyze theoretical and policy issues in international trade.

## II

A major shortcoming of Baldwin's use of the proportion of *R&D* workers in an industry is that it cannot discriminate between a theory of technological innovation as a stimulus to exports and a hypothesis relating trade patterns to the proportions of various labor skill groups. In order to find some unambiguous evidence supporting the technological innovation theory of

<sup>5</sup> The three alternative hypotheses are: 1) technological change as a determinant of trade flows, 2) economies of scale, and 3) differential endowments of various labor skill groups. These are discussed in Baldwin's article.

export determination, changes in the industries' production functions will be examined.

Discovery and implementation of new production processes is assumed to reduce cost, increase efficiency, and raise the productivity of the factor inputs. To isolate the changes in factor productivity from variation in the level and proportion of inputs, the industries are assumed to have the following production function:<sup>6</sup>

$$(2) \quad Q = Af(K, L)$$

Taking the time differential denoted by superscript  $t$  and performing the appropriate transformation yields

$$(3) \quad \frac{A^t}{A} = \frac{(Q/L)^t}{(Q/L)} - W_k \frac{(K/L)^t}{(K/L)}$$

which can be interpreted as: technical progress,  $A^t/A$ , equals the percent change in labor productivity after adjusting for changes in capital productivity and the proportion of capital to labor.

The data necessary to calculate Solow's measure of technical change ( $A^t/A$ ) for each industry were taken from the *Annual Survey of Manufacturing* with the change measured over the period 1961-67.<sup>7</sup> The U.S. trade embodiment of the Solow type of technical change can be determined by multiplying  $A^t/A$  by the export and import patterns. The ratio of the average technical change in exports to that of imports for 1967 is:

#### Solow Technical Change

U.S. Exports	20.5
U.S. Imports	18.2
Exports/Imports	1.13

<sup>6</sup> This is the formulation used by Robert Solow. It assumes the function is linear homogeneous and that technological change is neutral with respect to the inputs. Martin Beckmann and Ryuzo Sato show that this model best fits U.S. aggregate data when compared with eight other measures of technical change.

<sup>7</sup> The data on capital were poor for several reasons relating to the time period covered, and the measurement of depreciation and the share going to capital ( $W_k$ ). Alternative methods of calculating these variables were used, but the results were not sensitive to these variations.

TABLE 2—AVERAGE CHANGE IN PRODUCTIVITY EMBODIED IN 1967 EXPORTS

Country	Solow Technical Change Embodied in Exports
Netherlands	22.3
United States	20.5
Sweden	20.0
France	19.6
Germany	19.6
Japan	19.4
United Kingdom	19.3
Italy	19.0
Australia	18.4
Belgium-Luxembourg	17.9
Canada	16.3

Sources: Estimates of technological change from data in *Annual Survey of Manufactures*, export patterns from *Commodity Trade Statistics*, Series D.

This result is consistent with the hypothesis that a country will tend to export products for which cost saving innovations occur. However, it is not clear whether other countries are also exporting these products to the same extent. Using the estimates of technical change from U.S. data, but the trade patterns of other countries, the average productivity increases for exports have been calculated. Table 2 indicates that, except for the Netherlands, the U.S. exports goods reflecting a greater amount of technical change than any other major industrial country. The Netherlands' high ranking is due to large exports of chemicals, electrical machinery, office machinery, and petroleum production; all of which demonstrated relatively large technical change.

### III

One obvious deficiency which is common to most investigations of the determinants of trade patterns is the almost exclusive reliance on U.S. data concerning the economic characteristics of industries. The lack of comparable foreign data is the explanation for this omission, but in an effort to avoid a one-sided view of international trade this paper will attempt to introduce cross-country information at the aggregate and industry levels.

Table 2 presented cross-country compari-

TABLE 3—CHANGE IN EXPORT SHARE AND RATE OF GROWTH PER UNIT OUTPUT

Country	Growth of Output per Unit Input (1950-62)	Change in Export Share (1953-61)
Germany	4.43	66.0
Italy	4.25	76.5
France	3.65	6.5
Netherlands	2.79	20.0
Belgium	2.01	-3.4
U.S.	1.36	-27.5
U.K.	1.18	-21.7

Sources: Productivity changes from Denison, export shares see Table 2.

sons, but only U.S. data on productivity change was employed. Clearly, any export advantage resulting from an innovation will depend on the relative rates of innovation in all the countries. This information is not available on the industry level, but Edward Denison has calculated the rate of growth of output per unit input for seven countries on an aggregate basis, p. 140. Table 3 compares the change in the countries' aggregate export share, a measure of competitive strength, to the rate of growth of output per unit input for a similar time period. There is an obvious strong positive correlation between productivity increase and improvement in export performance.<sup>8</sup> This relationship provides further evidence supporting the technological change theory.

Cross-country data may also be applied to the hypothesis that a country with a large domestic market can better take advantage of economies of scale, and therefore export products which are produced by large operations. Joe Bain has estimated the proportion of workers in each industry who work in plants of less than 100 employees. He has done this for five major countries using data developed in each country. These percentages were then multiplied by the respective proportion of total exports each industry accounts for and summed to derive the proportion of total exports produced in small scale plants. Table 4 reports these figures along with data on gross national

TABLE 4—ECONOMIES OF SCALE IN EXPORTS AND MARKET SIZE

Country	Small Scale Embodied in Exports	GNP
U.S.	7.7	365.4
U.K.	18.2	47.5
France	19.1	30.7
Italy	29.3	19.1
Japan	30.4	19.0

Sources: Scale estimates from Bain, exports see Table 2; GNP figures from *International Financial Statistics*. Also, see fn. 9.

product which indicates the size of the domestic market.<sup>9</sup>

The United States has a much smaller proportion of its exports produced in small scale plants than the other four countries. The correlation between this scale-export measure and GNP is clearly negative which provides support for the hypothesis that large domestic markets encourage exports of large scale products.

However, it would be possible for the United States to export large scale products, and also to import this type of commodity. Therefore, a supplementary study was done to compare the scale content of exports to the scale content of imports for all eleven nations in the sample. For this purpose Hufbauer's estimated scale factor was used for all countries. The ratio of scale content of exports to imports for the United States was almost two, and as shown in Table 5 higher than the other nations. The correlation between the scale-trade ratio and market size as represented by total manufacturing output is +.77, and this adds further evidence for returns to scale as a determinant of trade patterns.

In conclusion, Baldwin's article has made a major contribution to understanding the determinants of trade patterns. This paper has demonstrated that by combining Baldwin's and Hufbauer's data, improved statistical results which conform to theoretical expectations are achieved. Furthermore, by

<sup>8</sup> The correlation between the two variables in Table 3 is +.91 and the rank correlation is +.89.

<sup>9</sup> Data on GNP is for 1954, and on scale; 1954 for U.S., 1951 for Italy, 1954 for U.K. and France, and 1956 for Japan.

TABLE 5—SCALE CONTENT OF EXPORTS AND IMPORTS AND MARKET SIZE

Country	Large Scale Content in Exports	Large Scale Content in Imports	Large Scale Content Exports/Imports	Manufacturing Output
Canada	439	464	.946	10.55
U.S.	596	306	1.947	173.04
Belgium	366	345	1.060	4.11
France	455	381	1.194	27.53
Germany	430	276	1.557	40.61
Italy	330	380	.868	17.40
Netherlands	456	358	1.293	5.55
Sweden	463	358	1.293	5.62
U.K.	242	224	1.080	32.22
Australia	313	495	.632	5.63
Japan	282	399	.706	21.56

Sources: Data for scale embodied in exports and imports and manufacturing output are from Hufbauer (1970) Tables 2, 3, and 4.

focusing on the process of technical change as indicated by the industry production function and by assembling cross-country evidence on economies of scale, additional independent evidence is presented to confirm several promising hypotheses of export determination.

#### REFERENCES

- J. S. Bain, *International Differences in Industrial Structure: Eight Nations in the 1950s*, New Haven 1966.
- R. E. Baldwin, "Determinants of the Commodity Structure of U.S. Trade," *Amer. Econ. Rev.*, Mar. 1971, 61, 126-46.
- M. Beckmann and R. Sato, "Aggregate Production Functions and Types of Technical Progress: A Statistical Analysis," *Amer. Econ. Rev.*, Mar. 1969, 59, 88-101.
- E. Denison, *Why Growth Rates Differ*, Washington 1967.
- W. Gruber, D. Mehta, and R. Vernon, "The R&D Factor in International Trade and International Investment of United States Industries," *J. Polit. Econ.*, Feb. 1967, 75, 20-37.
- G. C. Hufbauer, "The Impact of National Characteristics and Technology on the Commodity Composition of Trade in Manufactured Goods," in R. Vernon, ed., *The Technology Factor in International Trade*, New York 1970.
- D. Keesing, "The Impact of Research and Development on United States Trade," *J. Polit. Econ.*, Feb. 1967, 75, 38-48.
- W. W. Leontief, "Domestic Production and Foreign Trade: The American Capital Position Re-examined," *Proc. Amer. Philosophical Soc.*, Sept. 1953, 97, 331-49.
- R. M. Solow, "Technical Change and the Aggregate Production Function," *Rev. Econ. Statist.*, Aug. 1957, 39, 312-20.
- International Monetary Fund, *International Financial Statistics*, 17, no. 1, Jan. 1964.
- United Nations, *Statistical Papers Commodity Trade Statistics*, Series D, New York, 1953, 1960, 1961, 1967.
- , *Statistical Papers. Standard International Trade Classification, Revised*, Series M, no. 34, New York 1961.
- , *Yearbook of International Trade Statistics*, 1956 and 1964, New York 1957, 1965.
- U.S. Bureau of the Census, *Annual Survey of Manufactures*, Washington 1953, 1960, 1961, 1962, 1963, 1964, 1967.

# Determinants of the Commodity Structure of U.S. Trade: Reply

By ROBERT E. BALDWIN\*

Lawrence Weiser and Keith Jay have made an important contribution to our efforts to sort out the relative significance of innovational activities, different types of labor skills, and economies of scale as determinants of the commodity structure of international trade. In particular, they provide new evidence substantiating the importance of technological progress as a basis for the U.S. comparative advantage position as well as support for the view that scale economies should be given greater weight than my study indicated. However, there are still a number of questions and problems to be answered before very firm conclusions can be drawn on these matters.

In an effort to disentangle the relative importance of the innovational and skill factors, the authors calculate a measure of technical progress for various U.S. manufacturing industries between 1961 and 1967. However, their estimate of technical progress does not take explicit account of changes in the skill levels of industries during that period. Consequently, one still cannot be sure of the extent to which the measure represents innovational activity versus changes in the quality of labor.

Although my study indicates that the size of plants in export industries is considerably larger than in import-competing industries, the scale variable did not show up as significant in the various multiple regressions that included such variables as capital/labor ratios by industry, proportions of various skill groups, and R&D efforts. However, Weiser and Jay do obtain significant multiple regression results for the scale factor when they use Gary Hufbauer's measure of the scale variable and a different dependent variable than the one I used.

Their dependent variable is the ratio of U.S. exports of each industry to the industry exports of eleven leading exporting countries whereas my dependent variable was adjusted U.S. exports minus adjusted U.S. imports. It should be noted that when they used Hufbauer's scale variable and exports minus imports as the dependent variable, they failed to obtain a significant result for the scale variable. To a considerable extent, therefore, the issue seems to come down to what is the best dependent variable to use.

Clearly, if one is interested in the best indicator of export performance (as they are), a variable including imports is not appropriate. However, trade theory should, I think, generally focus on net trade flows rather than just exports or imports since the policy variables in which we usually are concerned are framed in net terms, for example, balance-of-trade or net employment effects of trade policy changes. As the authors are aware, it might well be that their scale variable would not show up as significant if they in some way also included the U.S. world import share by industry as part of their dependent variable. It would be interesting to extend their analysis of trade patterns along these lines as well as by taking account of skill changes in calculating a measure of technical progress.

## REFERENCES

- R. E. Baldwin, "Determinants of the Commodity Structure of U.S. Trade," *Amer. Econ. Rev.*, Mar. 1971, 61, 126-46.
- G. C. Hufbauer, "The Impact of National Characteristics and Technology on the Commodity Composition of Trade in Manufactured Goods," in R. Vernon, ed., *The Technology Factor in International Trade*, New York 1970.

\* Professor of economics, University of Wisconsin.

# Unions and Relative Real Wages

By MICHAEL J. BOSKIN\*

Much attention has been focused recently on the effects of unions on economic stability, resource allocation and income distribution. Almost always, the discussion begins with the effects of unions on labor earnings or wages.<sup>1</sup> Yet substantial agreement on the magnitude of the effect of unions on wages or earnings hardly seems close at hand. Among other studies on this subject, it is noted that the classic study by H. Gregg Lewis estimates a union/nonunion wage differential of about 10-15 percent in 1957-58; Leonard Weiss estimates about the same differential as Lewis; and Victor Fuchs, Frank Stafford, Adrian Throop, and Orley Ashenfelter and George Johnson estimate a much larger differential.

The question such studies should attempt to answer is whether and how much union membership increases wages facing individuals, holding constant other things such as education, race, sex, age, and occupation. The studies mentioned above are not entirely appropriate to answer this question. For example, some suffer from a potentially severe aggregation bias in examining *average* wages or earnings and the *percentage* of the labor force unionized and/or fail to disaggregate by race and sex. Those that attempt to examine opportunities facing individuals are forced to employ data on earnings rather than wages and thereby

build (at least partially) voluntary labor supply and demand decisions into their estimates. The purpose of this paper is to present new evidence on the relative *wages* of union and nonunion workers by applying recent advances in the hedonic method of price measurement to a new and rich source of data on individual workers.

In Section I, an equation relating wages to personal characteristics is developed which focuses on union membership and its interaction with race, sex, occupation, and geographical area. The equation extends work in this area by Robert Hall. In Section II, a brief discussion of the data is presented together with empirical estimates of the union/nonunion wage differential. Formal tests are made of some interesting hypotheses about the pattern of the relative wages of union and nonunion workers by race, sex, occupation, and geographical area. The results are in much closer accord with the estimates of Lewis and Weiss than those of Fuchs, Stafford, and Throop. In Section III, some concluding remarks are offered, including some observations on the limitations of this type of study.

## I. Estimation of a Real Wage Equation

The data used to estimate the real wage equation are taken from the 1967 Survey of Economic Opportunity (SEO). These data comprise an augmented version of the Current Population Survey, and they include all persons 14 years of age or over working at a wage or salary. The hourly wage rate is computed as an average over the week covered by the SEO. Wage rates are deflated by a geographical cost-of-living index derived from Bulletin 1570-5 of the U.S. Bureau of Labor Statistics as discussed in Boskin.

The basic equation is a regression of the real wage rate on a set of dummy variables representing personal characteristics. Separate regressions for each race-sex group were run, thereby allowing a complete set of

\* Department of economics, Stanford University. I wish to thank Robert Hall and Michael Hurd for valuable advice and assistance on earlier research from which this study stems, John Pencavel and the referee for advice, and Thomas Moore for programming assistance.

The technique used here to estimate wages was pioneered by Hall. The reader is referred to his excellent paper for a detailed discussion of this and other innovative techniques.

<sup>1</sup> All too frequently it ends there too; other possible influences of unions, for example, on nonpecuniary returns, etc., are usually ignored. A full-fledged theory of unionization is badly overdue, but I eschew any such attempt here as my task is the more modest one discussed above.

interactions between these and all other effects. In addition, I allow interactions between union membership and occupation, as well as between union membership and geographical area. Otherwise, all effects are assumed independent, for example, the occupation effect does not vary by educational attainment.

Table 1 presents the regression results both for the aggregate and disaggregate versions. The reference group, subsumed in the constant, in each case is 25–34 years old, nonunion full-time clerical workers, with no health effects on work capacity, twelve years of education, having resided in the United States at age 16 and currently living in an urban area in the western United States (white males with these characteristics for the aggregate regressions). Coefficients are the natural logarithms of the multiplicative effects of the associated characteristics and measure deviations *relative* to the reference group. Standard errors are in italics. Reference characteristics or characteristics for which insufficient data was available are replaced by a dash. An asterisk (\*) indicates the coefficients are statistically significant at the 5 percent level.

The equation performs well both in the aggregate and disaggregate forms. The standard error of the regression in each case is in the 35–40 percent or so range. Virtually all coefficients have the expected sign and are measured quite precisely. Even a casual glance at Table 1 reveals the striking variation in real wages across different population subgroups. The reader is referred to Hall for a discussion of the pattern of wages by age and education in the different race-sex groups. We focus instead in Section II on the new results reported here, namely, the pattern of wages classified by occupation, and the interaction of union membership with occupation and with location.

## II. Union/Nonunion Wage Differentials

This section presents evidence on the relative wage rates of union and nonunion workers, both overall and disaggregated by race and sex; we examine the pattern of union/nonunion wage differentials by oc-

cupation and geographic area, and the relative wages of white, black, male, female, union, and nonunion workers.

### *Occupation*

The effects of personal characteristics on wages may be seen most easily by converting the coefficients from Table 1 into the corresponding wage rates. Table 2 shows the occupational structure of wages and the interaction of occupation and union membership, other things held constant. The reference group remains 25–34 year old full-time workers with twelve years of education residing in an urban area in the western United States; the independence assumption implies that the estimated differentials are invariant with choice of the reference group.

For each occupation where the data permit, I have calculated a union/nonunion wage differential. The overall absolute differential ranges from practically zero to \$0.80 per hour for laborers, or \$1,600 per full-time year. In the craftsman, laborer, and operative occupations—areas of relatively strong unionizing activity—the relative wage differential runs from 15 to 25 percent. This amounts to approximately \$920, \$1,000, and the previously mentioned \$1,600 additional full-year potential earnings for union over nonunion members in the operative, craft and laborer groups. On the other hand, union service and sales workers and managers—areas of relatively light unionization—earn about the same wages as nonunion workers. Finally, we note that the estimated wage differential is comparable to Lewis' original estimate, but considerably smaller than those of the recent studies by Stafford, Throop, and Fuchs.<sup>2</sup>

The pattern for white males, of course, coincides with the overall pattern since they are the dominant group in the labor force. The pattern for black males is not very different. Unionized craftsman, operatives, and laborers all receive substantially higher wages than their nonunionized counterparts.

<sup>2</sup> Stafford also disaggregates by occupation, but his estimates of earnings differentials are larger than our estimate of *wage rate* differentials.

TABLE 1—REAL WAGE EQUATIONS<sup>a</sup>

Characteristic	Sex-Race Group				
	Male		Female		
	All	White	Black	White	Black
Constant	1.173* 0.012	1.101* 0.020	1.094* 0.027	0.855* 0.022	0.790* 0.023
Black	-0.113* 0.006	—	—	—	—
Female	-0.303* 0.007	—	—	—	—
Age 14-15	-0.572* 0.032	-0.652* 0.048	-0.456* 0.068	-0.263* 0.067	-0.638* 0.123
16-17	-0.315* 0.019	-0.364* 0.031	-0.253* 0.040	-0.149* 0.041	-0.220* 0.050
18-19	-0.267* 0.014	-0.336* 0.024	-0.163* 0.030	-0.237* 0.028	-0.170* 0.035
20-24	-0.119* 0.010	-0.162* 0.015	-0.078* 0.020	-0.089* 0.021	-0.076* 0.023
35-44	0.069* 0.008	0.106* 0.012	0.082* 0.016	0.014 0.019	0.022 0.018
45-54	0.083* 0.008	0.123* 0.012	0.081* 0.017	0.037* 0.019	0.055* 0.020
55-64	0.052* 0.010	0.099* 0.015	0.051* 0.020	0.012 0.022	-0.014 0.024
65+	-0.041* 0.019	-0.001 0.029	-0.065 0.040	-0.067 0.042	-0.022 0.043
Education 0-3 yr.	-0.250* 0.017	-0.294* 0.033	-0.182* 0.025	-0.116 0.066	-0.307* 0.039
4-6	-0.0194* 0.012	-0.212* 0.021	-0.150* 0.020	-0.188* 0.037	-0.172* 0.026
7-9	-0.128* 0.008	-0.156* 0.012	-0.089* 0.017	-0.079* 0.020	-0.126* 0.019
10-11	-0.063* 0.008	-0.065* 0.013	-0.066* 0.017	-0.039* 0.020	-0.070* 0.019
13-14	0.083* 0.010	0.075* 0.014	0.084* 0.026	0.078* 0.020	0.108* 0.025
15	0.075* 0.019	0.100* 0.027	0.061 0.044	0.071 0.038	0.003 0.054
16	0.254* 0.013	0.271* 0.018	0.204* 0.037	0.205* 0.029	0.260* 0.042
17-20	0.265* 0.016	0.262* 0.021	0.344* 0.048	0.223* 0.036	0.247* 0.052
Foreign residence at age 16	-0.035* 0.015	-0.082* 0.021	-0.006 0.041	-0.026 0.030	-0.043 0.049
Adverse health effect	-0.88* 0.009	-0.123* 0.014	-0.075* 0.019	-0.068* 0.023	-0.035 0.020
Usually work part-time	-0.140* 0.009	-0.290* 0.019	-0.196* 0.021	-0.106* 0.015	0.033* 0.016
Location: nonunion	-0.083* 0.010	-0.072* 0.016	-0.154* 0.025	-0.067* 0.020	-0.127* 0.030
Urban North-central	-0.079* 0.010	-0.051* 0.016	-0.153* 0.025	-0.086* 0.020	-0.047* 0.021
Urban South	-0.147* 0.009	-0.074* 0.015	-0.288* 0.020	-0.107* 0.020	-0.197* 0.017
Rural Northeast	-0.164* 0.026	-0.163* 0.034	-0.189 0.168	-0.097* 0.045	0.064 0.384
Rural North-central	-0.213* 0.018	-0.173* 0.025	-0.392* 0.188	-0.193* 0.031	-0.008 0.273
Rural South	-0.259* 0.011	-0.162* 0.018	-0.431* 0.025	-0.156* 0.024	-0.457* 0.024
Rural West	-0.112* 0.028	0.128* 0.038	-0.083 0.096	-0.087 0.050	0.148 0.124
Location: union	-0.123 0.017	-0.095* 0.023	-0.188* 0.033	-0.070 0.041	-0.148* 0.047
Urban Northeast	-0.046* 0.016	-0.035 0.022	-0.087* 0.029	0.027 0.044	-0.136* 0.045

Characteristic	Sex-Race Group				
	Male		Female		
	All	White	Black	White	Black
Urban South	-0.034 0.018	0.021 0.027	-0.133* 0.031	0.053 0.054	-0.155* 0.055
Urban West	0.073* 0.027	0.058* 0.010	0.022 0.017	0.030* 0.010	0.025 0.020
Rural Northeast	-0.105* 0.048	-0.086 0.056	-0.279 0.217	-0.012 0.114	-0.341 0.386
Rural North-central	-0.061 0.035	-0.032 0.038	0.084 0.374	-0.047 0.104	—
Rural South	-0.052 0.028	-0.001 0.035	-0.252* 0.062	0.011 0.083	—
Rural West	-0.063 0.061	-0.052 0.070	0.096 0.172	-0.196 0.182	0.395 0.547
Occupation: nonunion					
Professional/Technical	0.136* 0.013	0.161* 0.020	0.128* 0.036	0.133* 0.024	0.259* 0.037
Farmers (+managers)	-0.284* 0.041	-0.264* 0.053	-0.241* 0.066	-0.429 0.441	—
Managers, proprietors	0.148* 0.014	0.198* 0.020	0.053 0.050	0.055 0.032	0.100 0.070
Clerical	—	—	—	—	—
Sales	-0.155* 0.015	-0.010 0.023	-0.214* 0.058	-0.291* 0.025	-0.257* 0.052
Craftsman	0.033* 0.013	0.085* 0.019	0.020 0.027	-0.107 0.063	-0.108 0.077
Operatives	-0.137* 0.011	-0.077* 0.019	-0.117* 0.025	-0.141* 0.020	-0.173* 0.027
Private household	-0.516* 0.015	-0.364 0.154	-0.388* 0.154	-0.685* 0.036	-0.565* 0.024
Services	-0.267* 0.011	-0.172* 0.023	-0.250* 0.026	-0.328* 0.020	-0.266* 0.022
Farm laborer	-0.589* 0.021	-0.446* 0.034	-0.558* 0.035	-0.569* 0.128	-0.525* 0.060
Laborer	-0.236* 0.015	-0.155* 0.026	-0.168* 0.026	-0.259* 0.107	-0.316* 0.088
Occupation: union					
Professional/Technical	0.290* 0.039	0.326* 0.046	0.178 0.118	0.595* 0.130	-0.027 0.139
Farmers	—	—	—	—	—
Managers	0.101* 0.044	0.166* 0.050	-0.025 0.107	-0.034 0.168	—
Clerical	0.086* 0.035	0.073* 0.022	0.056* 0.023	-0.054 0.028	0.045 0.033
Sales	-0.129* 0.043	-0.045 0.063	-0.004 0.100	0.242* 0.082	-0.221 0.175
Craftsman	0.172* 0.018	0.217* 0.025	0.149* 0.036	0.014 0.148	0.090 0.116
Operatives	0.015 0.016	0.075* 0.025	0.038 0.032	-0.053 0.037	-0.031 0.040
Private household	-0.242 0.291	—	—	0.041 0.442	—
Service	-0.174* 0.025	-0.107* 0.050	-0.217* 0.043	-0.097 0.065	-0.176* 0.053
Farm laborer	-0.257 0.131	-0.105 0.233	-0.421* 0.173	—	-0.275 0.389
Laborers	0.036 0.023	0.088* 0.035	0.058 0.036	-0.183 0.182	0.158 0.142
Standard Error	0.41	0.40	0.37	0.43	0.38
Number of observations	24,627	9,949	4,850	5,908	3,920

<sup>a</sup> Standard errors in *italics*.

TABLE 2—RELATIVE REAL HOURLY WAGES OF UNION AND NONUNION WORKERS, BY OCCUPATION

Occupation		All	Male		Female	
			White	Black	White	Black
Professional/ Technical:	<i>NU</i>	\$3.70	\$3.53	\$3.39	\$2.69	\$3.05
	<i>U</i>	4.32	4.17	3.57	4.26	2.29
	%	19.0*	21.1*	5.9	66.1*	-32.2*
Farmers/Farm Managers:	<i>NU</i>	2.43	2.31	2.34	1.53	—
	<i>U</i>	—	—	—	—	—
	%	—	—	—	—	—
Manager/ Proprietor	<i>NU</i>	3.75	3.67	3.15	2.48	2.60
	<i>U</i>	3.58	3.55	2.91	2.27	—
	%	-5.3*	-3.9*	-7.8*	-8.9	—
Clerical	<i>NU</i>	3.23	3.01	2.98	2.35	2.20
	<i>U</i>	3.52	3.24	3.15	2.50	2.31
	%	9.1*	7.6*	5.7*	5.6*	4.6*
Sales	<i>NU</i>	2.77	2.98	2.41	1.76	1.82
	<i>U</i>	2.84	2.88	2.97	1.85	1.89
	%	2.3*	-3.3*	18.9*	3.7*	2.9
Craftsman:	<i>NU</i>	3.34	3.27	3.05	2.11	2.11
	<i>U</i>	3.84	3.73	3.46	2.38	2.57
	%	15.5*	15.5*	14.0*	11.1*	19.8*
Operatives:	<i>NU</i>	2.82	2.79	2.66	2.04	1.98
	<i>U</i>	3.28	3.24	3.10	2.23	2.28
	%	15.2*	15.1*	14.9*	8.0*	12.9*
Private Household:	<i>NU</i>	1.93	2.09	2.03	1.19	1.34
	<i>U</i>	—	—	—	—	—
	%	—	—	—	—	—
Service:	<i>NU</i>	2.48	2.53	2.33	1.69	1.80
	<i>U</i>	2.71	2.70	2.40	2.14	1.97
	%	7.4*	5.7*	2.5*	18.8*	7.2*
Farm Laborer:	<i>NU</i>	1.79	1.93	1.71	1.33	1.39
	<i>U</i>	—	—	—	—	—
	%	—	—	—	—	—
Laborer:	<i>NU</i>	2.55	2.53	2.52	1.82	1.72
	<i>U</i>	3.35	3.29	3.17	1.96	2.75
	%	24.7*	23.6*	21.5*	6.1	44.2*

Note: *NU*: nonunion workers.

*U*: union members.

%: percentage differential =  $w_u - (w_{NU})/w_{NU}$ .

\*: union and nonunion wages significantly different at the 5 percent level.

White females, however, do not appear to follow the pattern quite so closely. While unionized white females in the crafts enjoy wages 11 percent higher than nonunionized workers, the corresponding figures for opera-

tives and laborers are only 8.0 and 6.1 percent, respectively. Of course, these are occupations with a relatively small percentage of women. Black women demonstrate a pattern of union/nonunion relative

TABLE 3—RELATIVE REAL HOURLY WAGES OF UNION AND NONUNION WORKERS BY GEOGRAPHICAL AREA

Area		All	Male		Female	
			White	Black	White	Black
Urban Northeast:	<i>NU</i>	\$2.98	\$2.80	\$2.56	\$2.20	\$2.04
	<i>U</i>	2.86	2.72	2.47	2.19	2.03
	%	-3.7	-2.1	-2.9	-0.3	-0.4
Urban Northcentral:	<i>NU</i>	2.99	2.86	2.56	2.16	2.24
	<i>U</i>	3.09	2.91	2.74	2.42	2.05
	%	3.1	1.6	5.8	11.1	-8.1
Urban South:	<i>NU</i>	2.79	2.79	2.25	2.11	1.93
	<i>U</i>	3.13	3.07	2.61	2.48	2.10
	%	10.4	9.3	12.5	15.6	7.0
Urban West:	<i>NU</i>	3.23	3.01	2.98	2.35	2.20
	<i>U</i>	3.47	3.19	3.04	2.42	2.26
	%	7.5	6.0	2.2	3.1	2.5
Rural Northeast:	<i>NU</i>	2.75	2.56	2.93	2.14	2.51
	<i>U</i>	2.91	2.76	2.33	2.32	1.68
	%	5.2	6.8	-20.1	8.0	-35.5
Rural Northcentral:	<i>NU</i>	2.61	2.53	2.02	1.94	2.33
	<i>U</i>	3.04	2.91	3.25	2.38	—
	%	13.3	12.6	41.2	12.9	—
Rural South:	<i>NU</i>	2.50	2.56	1.94	2.01	2.73
	<i>U</i>	3.07	3.01	2.32	2.38	—
	%	17.8	14.8	12.7	15.5	—
Rural West:	<i>NU</i>	2.89	2.65	2.75	2.16	2.73
	<i>U</i>	3.04	2.86	3.29	1.93	3.49
	%	4.5	6.9	19.1	-9.5	32.4

Notes: See Table 2.

wages closer to those of white and black males than to white women for the three occupations just mentioned.

For the other occupations, the data are more difficult to interpret. Certainly there exist wide variations in the union/nonunion differential by race and sex for some occupations. For example, the differential in wages is only about 6 percent for black male professionals compared to over 21 percent for white males.

Yet not too much should be read into this result. The definitions of the occupations are too broad to allow us confidently to speak as if blacks and whites, or men and women, were necessarily employed at the same jobs. Also, we have not allowed oc-

cupation-location interaction. Since the relative unionization of whites and blacks differs by location, we must keep this in mind in interpreting the results.

Finally, the assumption of independence of the union and occupation effects usually yields larger estimates of the differential than a weighted average of the separate occupation estimates. For example, the overall differential for black males, 23.1 percent, is larger than *any* of the separate occupation differentials. We infer that the interactions are indeed present.

#### Geographical Area

The geographical pattern of the union/nonunion wage differential—again making

the restrictive assumption it is independent of occupation—is also worth noting, and is shown in Table 3. The overall differential is largest in the South, about 10 percent for urban areas and 18 percent for rural, and smallest in the Northeast where the rural differential is only 5 percent and does not differ significantly from zero for urban areas. Again, the pattern of the differential by race and sex differs somewhat across regions. For example, for black males it is 41 percent in the rural Northcentral compared to 13 percent for white males. The data reject the hypothesis of no difference between the races and between the sexes in (the geographic and occupational) union effects.<sup>3</sup>

#### Race and Sex Discrimination

It is interesting to note not only the wage differential between union and nonunion workers by race and sex, but also the relative wage differential between sexes for union and nonunion workers. The data in Table 2 suffice for this purpose; we calculate the black/white and female/male relative wage differential between union and nonunion workers.<sup>4</sup> Table 4 presents these results for those occupations with a sufficient number of union members to make the calculation meaningful. Again, the pattern varies somewhat by occupation, race, and sex. Most notable is the apparent lack of a differential between union and nonunion relative wages

<sup>3</sup> The tabulation below presents evidence in the form of formal tests of the hypotheses that the union effects do not vary by race and by sex. These hypotheses are rejected for each sex group for the test of no difference between the races, and for each race for the test of no differences between the sexes.

Test of:	Group	Critical	
		F-Statistic	F, 5 Percent level
No difference in Races in Union Effects	Males	4.9	1.5
	Females	5.8	1.5
No difference in Sexes in Union Effects	Whites	7.4	1.5
	Blacks	4.2	1.5

<sup>4</sup> The possibility of excluded variables suggests caution in attributing the differentials solely to discrimination.

TABLE 4—RELATIVE WAGE DIFFERENTIAL OF BLACK AND WHITE, FEMALE AND MALE, UNION AND NONUNION WORKERS (in percent)

Occupation	Race <sup>a</sup>		Sex <sup>b</sup>	
	Male	Female	White	Black
Managers	-4	—	-5	—
Clerical	-2	-1	-1	-1
Sales	+27	-1	+5	-15
Craftsman	0	+8	-1	+7
Operatives	0	+5	+24	+42
Service	-3	-13	+19	+6
Laborer	-1	+48	-15	+27

<sup>a</sup> The appropriate ratio is

$$\left[ \left( \frac{w_N}{w_W} \right)^U - \left( \frac{w_N}{w_W} \right)^{NU} \right] / \left( \frac{w_N}{w_W} \right)^{NU}$$

<sup>b</sup> The appropriate ratio is

$$\left[ \left( \frac{w_F}{w_M} \right)^U - \left( \frac{w_F}{w_M} \right)^{NU} \right] / \left( \frac{w_F}{w_M} \right)^{UN}$$

of black and white males. It appears that racial discrimination *within* unions is no worse for black males than it is outside of unions. If, however, my results from non-union/union wage differentials for black males may be taken as a rough guide to the discriminatory effects of *exclusion* from union membership, a different conclusion emerges. For other groups, we note that black female union laborers do considerably better relative to white females than their nonunion counterparts and that the evidence on relative discrimination (loosely defined) against women points to an improved economic position of union females relative to nonunion females for black and white operatives and black laborers.

### III. Conclusion

We have examined a new and rich source of data to determine the existence and magnitude of a union/nonunion wage differential. Our general conclusion has been that the relative wages of union and nonunion workers varies by race, sex, location, and occupation. The differential was smaller than most recent estimates and most pro-

nounced for craftsman, operatives, and laborers. We have discussed the need for the debate on the effects of unions to focus on wages rather than earnings and on individual rather than industry averages.

We must mention two points concerning the comparability of this study with previous investigations in this area. First, we examine data for 1967, a year of relatively low unemployment. Our generally smaller estimates of the union/nonunion wage differential may be due in part to cyclical influences on the wage differential. It may also be due in part to the relatively complete disaggregation attempted here.<sup>5</sup>

#### REFERENCES

- O. Ashenfelter and G. Johnson, "Unionism, Relative Wages, and Labor Quality in U.S. Manufacturing Industries," unpublished mimeo 1969.
- G. Becker, *The Economics of Discrimination*, Chicago 1957.
- <sup>5</sup> Some previous researchers have used the same degree of disaggregation as that reported here; however, disaggregate data on wage rates has in general been unavailable, and these researchers had to focus instead on earnings.
- M. Boskin, "The Economics of the Labor Supply," in G. Cain and H. Watts, eds., *Labor Supply and Income Maintenance*, Markham 1972, forthcoming.
- V. Fuchs, *The Service Economy*, New York 1968.
- R. Hall, "Wages, Income and Hours of Work in the U.S. Labor Force," in G. Cain and H. Watts, eds., *Labor Supply and Income Maintenance*, Markham 1972, forthcoming.
- H. G. Lewis, *Unionism and Relative Wages in the United States*, Chicago 1963.
- F. Stafford, "Concentration and Labor Earnings: Comment," *Amer. Econ. Rev.*, Mar. 1968, 58, 174-80.
- A. Throop, "The Union-Nonunion Wage Differential and Cost Push Inflation," *Amer. Econ. Rev.*, Mar. 1968, 58, 79-99.
- L. Weiss, "Concentration and Labor Earnings," *Amer. Econ. Rev.*, Mar. 1966, 56, 96-117.
- U.S. Bureau of the Census, "Survey of Economic Opportunity, 1967," Washington 1967.
- U.S. Bureau of Labor Statistics, *Three Standards of Living, Spring 1967*, Bulletin 1570-5, Washington 1968.

# Commodity Price Equalization: A Note on Factor Mobility and Trade

By FRANK FLATTERS\*

In his paper "International Trade and Factor Mobility," Robert Mundell showed that under the usual assumptions of the factor price equalization theorem, impediments to commodity movements will stimulate factor movements and *vice versa*. Moreover, in a two-country, two-commodity, two-factor model, the movement of a single factor will lead to the equalization of factor and commodity prices. The two countries will also consume the same amount of both commodities as under free trade; further, the factor importing country will produce the same amount of its "import good" and the factor exporting country the same amount of its "export good." The factor importing country will, however, produce an additional amount of its export good just sufficient to pay for the services of the imported factor, and the factor exporting country will produce correspondingly less of its import good.

A difficulty with this analysis, as Mundell himself points out, p. 106, is that although it is impediments to commodity trade which give rise to the factor movements from one country to the other, it must be assumed that goods sent to the factor exporting country in payment for factor services are not subject to such impediments. Otherwise there will remain a wedge between domestic and foreign prices of the mobile factor and neither commodity nor factor prices will be equalized.<sup>1</sup> Accordingly, it is only under very

special circumstances that factor mobility will be in fact a perfect substitute for commodity trade in achieving complete equalization of factor and commodity prices.

It seems to me that Mundell's problem arises only because he did not follow through completely the logic of his own argument. In attempting to show that factor mobility can be a substitute for commodity trade when the latter is subject to impediments, Mundell allows only one factor to move and necessarily concludes that in equilibrium commodities must still be exchanged to pay for the imported factor services. I shall demonstrate in this note that under the assumptions made by Mundell there will be no need for commodity trade to pay for factor services if both factors are free to move. There is indeed no reason to restrict mobility to one factor only, and the more general formulation suggested here is simply the converse of the factor price equalization theorem where *both* commodities and not just one are free to move.<sup>2</sup>

also suggests: "This problem could have been avoided had we allowed the capitalist to consume his returns in the country where his capital was invested." However, he fails to point out that the equilibrium production position in *A* can no longer be any arbitrary point beyond *P'* along the Rybczynski line (*R<sub>K</sub>* in Figure 1). Rather, equilibrium will be uniquely determined at that point along the Rybczynski line where the amounts of cotton and steel produced in excess of *S* will be exactly equal to the amounts demanded by the migrant capitalists at international prices. This is the only equilibrium point consistent with no trade in cotton and steel.

<sup>2</sup> Ernest Nadel used a model which is virtually identical to the one presented here (with both factors mobile) in order to examine the question of the optimality of various trade and factor flows when there are impediments to both. In so doing, however, he fell prey to the same difficulty that Mundell faced. Since in Nadel's model all trade and factor flows are subject to impediments, no combination of commodity and/or factor movements can equalize world prices and hence the set of equilibria which he considers cannot be achieved.

\* Assistant professor of economics, Queen's University. This paper was completed while I was in residence as a graduate student at the Johns Hopkins University. I am indebted to Bela Balassa for his encouragement and many useful suggestions on successive drafts of this paper. Trent Bertrand, James Melvin, Ernest Nadel, and Douglas Purvis also provided helpful comments at various stages.

<sup>1</sup> Mundell attempts to dodge this problem by relabeling the impediments to factor payments as imperfections in capital mobility, p. 106. On the same page, he

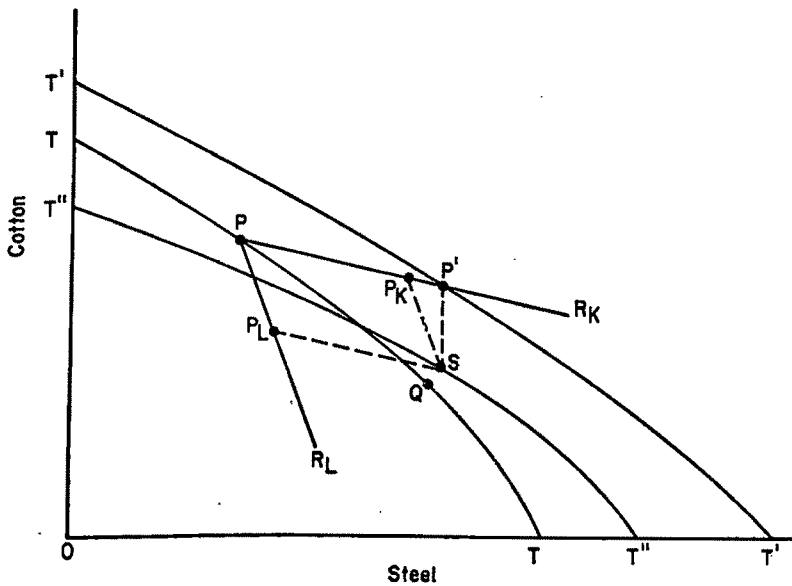


FIGURE 1

The demonstration of this result can be somewhat brief, since most of the arguments are to be found already in Mundell's paper. Following Mundell it will be assumed that country *A* is relatively labor abundant and *B* is capital abundant; in turn, the production of cotton is labor intensive relative to that of steel. It is also assumed initially that payment for factor services is received and spent in the country where the factor originates. The implications of relaxing this assumption will be shown later. Country *A* is assumed to be a small country relative to *B*.<sup>3</sup>

In autarky *A*'s production possibilities are represented by the transformation curve  $TT$  in Figure 1 or the efficiency locus  $00'$  in Figure 2, and the autarky production and consumption point is given by  $Q$  in both diagrams. With the introduction of free commodity trade at a given world price ratio *A* produces at  $P$  (in both figures) and consumes at  $S$  (in Figure 1), and its commodity and factor prices are identical to those existing in the world market.

Now suppose that international com-

modity movements are subject to impediments.<sup>4</sup> Mundell analyzed the case where capital was perfectly mobile between the two countries and showed that the equilibrium consumption point in country *A* is at  $S$ , the same as under commodity trade. This is because capital will continue to move into country *A* until factor prices are equalized. Since relative factor prices as well as the domestically owned factor bundle and thus the level and distribution of domestic income are then the same as under free trade, product prices will also be identical and the same commodity bundle will be consumed. However, since there has been an inflow of capital services into country *A* (measured by the distance  $PP'$  in Figure 1 and  $KK'$  in Figure 2), its new production possibilities are represented by  $T'T'$  and  $00''$ , and the new production point will be  $P'$ . Correspondingly, the distance  $P'S$  in Figure 1 represents the payment for the imported capital services. Thus, while Mundell's purpose has

<sup>3</sup> As Mundell showed, country size does not affect any of the results of this model.

<sup>4</sup> For the purposes of this paper it is sufficient to leave the exact type of impediment undefined. However it should be noted that different types and structures of trade impediments do in fact lead to quite different results. These differences are analyzed at some length as part of my doctoral dissertation.

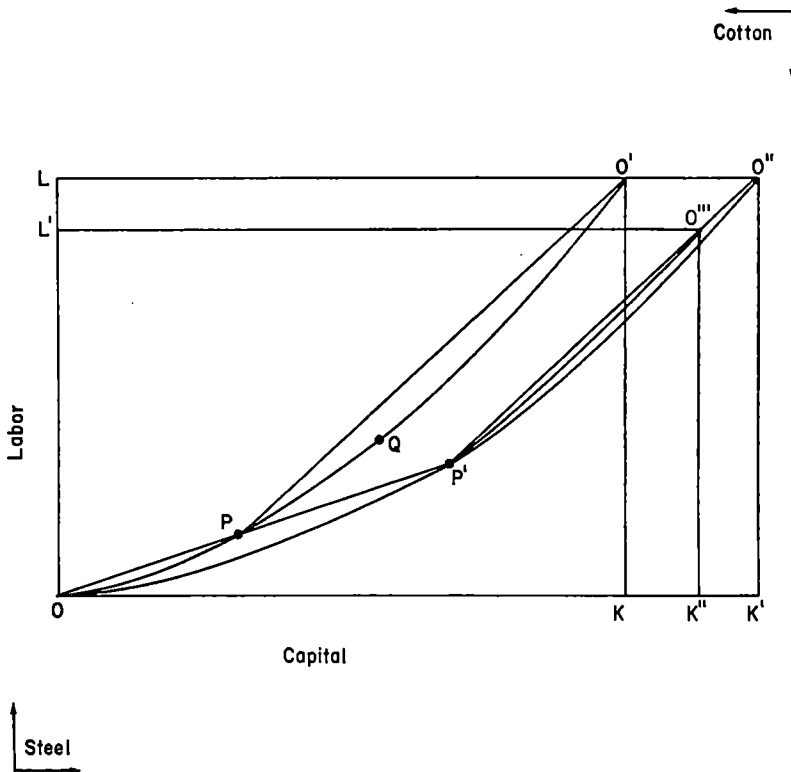


FIGURE 2

been to illustrate a situation where factor movements are a substitute for commodity trade, his results involve the trade of commodities for factor services.

What happens now if it is assumed that both factors are mobile? With impediments to commodity trade, country *A* will begin to import capital services, as in the Mundell case, but will now also export labor services. This movement of factor services will continue until relative factor prices are equalized. Due to the one-to-one correspondence between relative factor prices and relative commodity prices in each of the two countries under the assumptions of the factor equalization theorem, the movement of factors will also result in the equalization of commodity prices and hence even a small impediment to commodity trade will exclude the possibility of such trade. With the same domestically owned factor bundle the consumption point will once again be at *S*, the same as under free trade. But while in

the Mundell case it was shown that any movement of capital beyond a certain minimum amount which would move production at least as far as *P'* along the Rybczynski line ( $R_K$ ) would be consistent with equilibrium, in the present case there is only one possible exchange of factors.<sup>6</sup> Since movements of commodities are impeded by our initial assumption, *A* must import just sufficient capital ( $PP_K$  in Figure 1) and export just sufficient labor ( $PP_L$ ) to produce at its consumption point *S*, and the value of these two bundles of factor services will be equal at the world market factor prices. In the new equilibrium *A*'s production possibilities are represented by the curve  $T''T''$  and  $OO''$ , respectively, and the production point

<sup>6</sup> This is true since under the assumption of linear homogeneous production functions factor intensities are uniquely determined by factor and product prices and hence in equilibrium only one set of factor intensities can be chosen, which is possible with only one set of factor flows.

chosen is represented by  $S$  in Figure 1 and  $P'$  in Figure 2 (reading from the origin  $O'''$  instead of  $O''$ ). In the new equilibrium country  $A$  is exchanging  $PP_L$  (or  $L'L$  in Figure 2) of labor services for  $PP_K$  (or  $KK''$ ) of capital services rather than trading  $P'S$  of cotton for  $PP'$  (or  $KK'$ ) of capital services as in Mundell. Thus a point of equilibrium is reached where it is not necessary to contradict the initial assumption of impediments to commodity movements and it has been shown that factor mobility is a complete substitute for free commodity trade in equalizing factor and commodity prices.

It remains now to show what happens if factor owners are allowed to migrate when they sell their factor services abroad; that is, if it is assumed that they receive and spend their incomes in the country where their services are sold.<sup>6</sup> It can be expected that in general a different consumption point will be reached than in the previous case except when the additional assumption is made that all capitalists and labor owners have identical and homothetic preferences.<sup>7</sup> Since it has been shown already that the incomes of labor owners from  $A$  who sold their services in country  $B$  were equal to the incomes of the capital owners from  $B$  who sold their services in  $A$ , the aggregate consumption expenditures will be the same in both countries in the two cases. And if capitalists and laborers have identical preferences, the composition of the consumption bundles will also be the same in each case in both countries, and so the equilibrium in the case of no migration will be still consistent with equilibrium under this assumption. If the preferences of the migrating laborers differ from those of the migrating

capitalists, then each country's consumption point would be different than under free commodity trade, but total world consumption of both commodities would remain the same.

It can be seen then that with a generalization of Mundell's model, it has been possible to overcome the basic difficulty of that paper by showing that the equalization of factor and commodity prices through factor movements can be achieved without commodity trade in payment for factor services. In the process a theorem on commodity price equalization parallel to the factor price equalization theorem has been obtained: under identical assumptions commodity prices will be equalized through factor movements just as factor prices are equalized through commodity trade. Expressed differently, the equalization of factor and commodity prices can be attained through either the free movement of factors or the free exchange of commodities.<sup>8</sup> More generally, this model can be regarded as a mirror image of the traditional Heckscher-Ohlin analysis of commodity trade.

#### REFERENCES

- J. S. Chipman, "International Trade and Factor Mobility: A Substitution Theorem," in J. N. Bhagwati et al., eds., *Trade, The Balance of Payments and Growth*, Amsterdam 1971.
- F. Flatters, "Essays on Factor Mobility and Trade," doctoral dissertation in progress, Johns Hopkins Univ.
- R. A. Mundell, "International Trade and Factor Mobility," *Amer. Econ. Rev.*, June 1957, 47, 321-35; reprinted in R. Caves and H. G. Johnson, eds., *Readings in International Economics*, Homewood 1968, 101-14.
- E. Nadel, "Capital Goods, Intermediate Goods, and the Volume of Trade," *Can. J. Econ.*, May 1971, 4, 238-49.
- D. D. Purvis, "Technology, Trade and Factor Mobility," *Econ. J.*, forthcoming.

<sup>6</sup> If, with both factors mobile, only one factor migrates while the other consumes its income in its country of origin, the solution that will obtain will be that discussed in fn. 1, with only the migrating factor moving. Any other solution would necessitate a flow of goods and hence a difference between domestic and foreign prices due to the assumed impediments to commodity movements.

<sup>7</sup> This is not an unreasonable assumption to make if we wish to concentrate on the strictest version of a Heckscher-Ohlin type model and abstract from everything except the production effects of differences in factor endowments to explain patterns of international specialization.

<sup>8</sup> Note that this theorem depends critically on all the assumptions of the standard theorem on factor price equalization. Recent papers by John Chipman and Douglas Purvis have examined the degree to which factor mobility and trade are substitutes (or complements) when the assumptions of incomplete specialization and identical technologies are violated.

# A Theory and Test of Credit Rationing: Some Generalizations

By VERNON L. SMITH\*

This comment suggests certain extensions and generalizations of the theory of credit rationing discussed in the stimulating recent paper by Dwight Jaffee and Franco Modigliani (J-M). The J-M model derives a bank's optimal loan offer curve on the assumption (contained in the earlier literature cited)<sup>1</sup> that the size of the borrower-firm's investment is fixed and therefore the random outcome of the investment is independent of the size of the loan. It will be shown that a relaxation of these assumptions lays bare a fundamental condition of dependence—an "externality"—between a lender and a borrower under risk. The important consequence of this is to permit a deeper understanding of the debtor-creditor contract. An inevitable consequence of risky investment by a borrower is to cause the lender's terminal fortunes to depend on those of the borrower, and in particular to depend on the borrower's equity in the investment. This equity acts as an external economy to the lender.

## I. Investor Demand for Debt Finance

Consider an individual faced with an uncertain investment project. He will invest  $y$  in the project (his equity in a wholly owned firm), and intends to supplement this equity by obtaining a loan in the amount  $z$  to the firm at a contractual interest rate  $r^*$ . The total capitalization of the firm is thus  $y+z$  (equity+debt). The gross cash flow or proceeds of the investment, at the end of a single period, is a random variable  $X$ . Under stochastic constant returns (see Quirk, p. 213), the rate of return per dollar invested,  $\theta$ , is distributed independently of the amount invested. Hence,

\* University of Massachusetts.

<sup>1</sup> The contribution of James Quirk does not make this assumption. Quirk seems to have provided the first interpretation of credit rationing in terms of a "... demand curve (for a firm's notes) ... of unit elasticity" (p. 224).

$$(1) \quad X = (y + z)(1 + \theta),$$

where  $\theta$  has a subjective probability density  $G'(\theta) = g(\theta)$ . The variable  $X$  in our model corresponds to "the firm's end of period value" (p. 851) in J-M.<sup>2</sup> A formulation such as in equation (1) (or equation (1') of the footnote), is essential to an analysis dealing with the choice between equity and debt financing, and with the character of the debtor-creditor relationship.

In addition to the risky project, the investor is assumed to have available a secure investment opportunity such as short-term government bills paying a guaranteed return  $r < r^*$ . In our simple model,  $r$  is the opportunity cost of funds to the only available risky investment project. It is further assumed that the investor has initial wealth  $W_0 = u + y$  to be allocated between investment in riskless bonds,  $u$ , and equity investment  $y$ , in his firm. We assume that the investor's risky investment activity (his firm) is incorporated, with limited liability, so that the investor's personal account purchases of riskless bonds are protected from lender claims. Hence, if the investor's strictly concave utility function for terminal wealth is  $U(W)$ , his expected utility can be written

$$\begin{aligned} B = & U[(W_0 - y)(1 + r)]G(\theta^*) \\ (2) \quad & + \int_{\theta^*}^{\infty} U\{(W_0 - y)(1 + r) \\ & + (1 + \theta)(y + z) - (1 + r^*)z\} dG(\theta), \end{aligned}$$

where  $\theta^* = [(1 + r^*)z / (y + z)] - 1$  is the rate of return below which the firm is in default on its loan. In the event of default, the bank

<sup>2</sup> More generally, one would assume

$$(1') \quad X = \phi(y + z, s)$$

where  $s$  is a random state variable whose probability distribution induces a distribution on  $X$  given  $y + z$ , and  $\partial\phi/\partial(y + z)$  is the state-contingent marginal product of capital to the firm.

receives all of the cash flow  $(1+\theta)(y+z)$ , and the investor receives nothing on his equity investment and ends the period with only his buffer investment in riskless bonds. If  $\theta > \theta^*$ , the firm receives all the cash flow in excess of contractual bond payments, i.e.,  $(1+\theta)(y+z) - (1+r^*)z$ , and the investor ends the period with a return on his equity as well as his riskless bond investment.<sup>3</sup> If our investor chooses  $y$  and  $z$  so as to maximize  $B$ , we have, for an interior solution, the first-order partial derivative conditions,  $B_y = 0$ ,  $B_z = 0$ . These conditions are assumed to define (from the Jacobian implicit function theorem) demand functions for equity and loan capital, say<sup>4</sup>

$$(3) \quad y = Y(r^*, r, W_0)$$

$$(4) \quad z = Z(r^*, r, W_0)$$

<sup>3</sup> To see how (2) is generalized under the productivity hypothesis (1'), let  $\phi$  have an inverse  $s = \psi(X, y+z)$ , and assume  $\phi_s > 0$  for all  $(y+z, s)$ . If  $s$  has the probability function  $H(s)$ , and

$$s^* = \psi[X^*, y+z] = \psi[(1+r^*)z, y+z],$$

then

$$\begin{aligned} B &= U[(W_0 - y)(1+r)]H(s^*) \\ (2') \quad &+ \int_{s^*}^{\infty} U\{(W_0 - y)(1+r) + \phi(y+z, s) \\ &- (1+r^*)z\} dH(s) \end{aligned}$$

By substitution, this calculation reduces to (2) when  $X = \phi(y+z, s) \equiv (y+z)\phi(s) = (y+z)(1+\theta)$ , for we then have  $s = \psi(X, y+z) \equiv \phi^{-1}(1+\theta)$ ,  $dG(\theta) = dH(s)$ , and  $X^* = (1+r^*)z = (y+z)\phi(s^*) = (y+z)(1+\theta^*)$ .

<sup>4</sup> In (2) we assume  $B = B(y, z | r, r^*, W_0)$  is concave in  $(y, z)$ , i.e.,  $B_{yy} < 0$ ,  $B_{zz} < 0$ ,  $B_{yy}B_{zz} - B_{yz}^2 > 0$ . Concave  $U(W)$  is necessary but not sufficient for concave  $B$ , as can be seen by computing

$$\begin{aligned} B_{yy} &= (1+r)^2 U''[\cdot] G(\theta^*) \\ &+ \int_{\theta^*}^{\infty} U''[\cdot] \{(\theta - r)^2 dG(\theta) \\ &+ (1+\theta^*) U'[\cdot] g(\theta^*) (1+r^*)z/(y+z)^2\} \\ B_{zz} &= \int_{\theta^*}^{\infty} U''[\cdot] \{(\theta - r^*)^2 dG(\theta) \\ &+ (1+\theta^*) U'[\cdot] g(\theta^*) y^2 (1+r^*)/z(y+z)^2\} \\ B_{yz} &= \int_{\theta^*}^{\infty} U''[\cdot] \{(\theta - r)(\theta - r^*) dG(\theta) \\ &- (1+\theta^*) U'[\cdot] g(\theta^*) y(1+r^*)/(y+z)^2\} \end{aligned}$$

If  $U(W)$  is linear in wealth, we have  $B_{yy} > 0$ ,  $B_{zz} > 0$ ,  $B_{yy}B_{zz} = B_{yz}^2$  and  $B$  is not concave in  $(y, z)$ .

These equations assume that the lender does not ration capital so that the borrower has no difficulty obtaining the loan  $z$  which is optimal at interest  $r^*$ .

## II. Lender Supply of Debt Finance

We now consider a banker or other lender, with initial wealth  $w_0$ , and confronted with only two alternatives. He can buy government bonds paying the sure return  $r$ , which is his opportunity cost of lending at risk, and he can lend for the same period of time to the above investor at the risky rate  $r^*$ . The lender has a subjective density  $F'(\theta) = f(\theta)$  which is in general different from  $g(\theta)$ , and concave utility function  $V(w)$ , where  $w$  is terminal wealth. Hence, our banker will want to choose the amount of the loan,  $z$ , and the amount of riskless bonds  $(w_0 - z)$ , so as to maximize

$$\begin{aligned} L &= V[(w_0 - z)(1+r)]F(-1) \\ &+ \int_{-1}^{\theta^*} V\{(w_0 - z)(1+r) \\ (5) \quad &+ (z+y)(1+\theta)\} dF(\theta) \\ &+ V[(w_0 - z)(1+r) + z(1+r^*)] \\ &\cdot [1 - F(\theta^*)], \quad 0 \leq z \leq w_0 \end{aligned}$$

The first term in (5) is the banker's expected terminal utility in the event the investor is bankrupt, with the full amount of the loan lost.<sup>5</sup> The second term is the banker's expected terminal utility in the event of partial default on the loan, and the third term is expected utility in the event of full repayment plus contractual interest.

For maximum  $L$ , the first-order condition is<sup>6</sup>

<sup>5</sup> Referring to fnn. 2 and 3, assume  $X=0$ , for all  $s \leq s_0$ . Thus if  $K(s)$  is the lender's probability function on  $s$ , we have

$$\begin{aligned} L &= V[\cdot]K(s_0) \\ (5') \quad &+ \int_{s_0}^{\infty} V\{(w_0 - z)(1+r) + \phi(y+z, s)\} \\ &\cdot dK(s) + V(\cdot)[1 - K(s^*)] \end{aligned}$$

for expected utility under nonconstant returns to scale.

<sup>6</sup> We also have

$$\partial^2 L / \partial z^2 = (1+r)^2 V''[\cdot] F(-1)$$

$$(6) \quad (\partial L / \partial z^0) = - (1 + r) V'[\cdot] F(-1) \\ + \int_{-1}^{\theta^*} V''\{\cdot\} (\theta - r) dF(\theta) \\ + (r^* - r) V'(\cdot) [1 - F(\theta^*)] \stackrel{<}{\geq} 0,$$

where if  $<$  holds,  $z^0 = 0$ ; if  $>$  holds  $z^0 = w_0$ .

It will be optimal for the lender to refuse to make a positive loan if  $(\partial L / \partial z^0)_{z^0=0} \leq 0$ . Carrying out the calculations for  $z^0 = 0$ , we have<sup>7</sup>

$$(7) \quad r^* \leq [r + F(-1)] / [1 - F(-1)]$$

This reduces to the J-M condition,  $r^* \leq r$  (also see Smith), if the chance of bankrupt default,  $F(-1)$ , is zero.

Consequently, equation (6) defines implicitly the banker's supply function of debt finance,  $z = S(r^*; r, y)$ . If we let  $r^* = [r + F(-1)] / [1 - F(-1)]$  be that risky rate of interest below which the banker is unwilling to make a loan, then the supply curve in the plane  $(r^*, z)$  has intercept  $z = 0$ , at  $r^* = r^*$ .<sup>8</sup>

But the most significant socioeconomic property of equations (5)–(6), and therefore the supply function  $S$ , is the fact that they must inevitably contain the borrower's equity decision variable,  $y$ , as a parameter. No rational lender can make a loan decision without knowledge of the investor's equity decision. This model provides an explanation of why bankers not only want to know how

investors are going to use the proceeds of a loan (the banker must estimate  $F(\theta)$ ), but also why bankers want to know such financial state variables as the investor's equity. Indeed, the almost universal practice of relating loan size to equity in mortgage, consumer, and security loans is a manifestation of lender dependence on the borrower. This externality is unilateral, however. That is, in equation (4) we see that the borrower's loan demand function depends only on the trading terms and his wealth  $(r^*, r, W_0)$ , and is independent of the financial state of the banker. This explains why debtor-creditor negotiations appear rather superficially to be one-sided. What borrower ever delves into the financial affairs of a banker, except perhaps to ask if he has the money for a loan? But to a banker, the borrower's financial affairs are of the *essence* of their contractual arrangement. All these considerations are important in understanding the nature of the loan contract and why it differs from an ordinary commodity exchange.

### III. Effect of Equity on Lender Supply of Finance

Lender supply behavior under expected utility maximization has several special characteristics which it is of interest to verify.

**PROPOSITION 1.** *The borrower's equity in an investment is an external economy to the lender.*

This is proved by evaluating  $\partial L / \partial y$  from (5), and noting that

$$\partial L / \partial y = \int_{-1}^{\theta^*} V''\{\cdot\} (1 + \theta) dF(\theta) > 0$$

By differentiating (5') of footnote 5, the result can be verified under nonconstant returns.

**PROPOSITION 2.** *Under constant returns and with utility linear in wealth, the borrower's equity is an external economy at the lender's loan margin, i.e.,*

$$\frac{\partial^2 L}{\partial y \partial z} > 0$$

$$+ \int_{-1}^{\theta^*} V''\{\cdot\} (\theta - r) dF(\theta) \\ + (r^* - r) V''(\cdot) [1 - F(\theta^*)] \\ + V'(\cdot) f(\theta^*) (\theta^* - r^*) \\ \cdot y(1 + r^*) / (y + z)^2 < 0$$

for all  $z$ , if  $V'' \leq 0$ , since  $\theta^* < r^*$  (note that  $\lim_{z \rightarrow \infty} \theta^* = r^*$ ).

<sup>7</sup> This result holds also under nonconstant returns. Differentiating (5') and evaluating  $(\partial L / \partial z^0)_{z^0=0} \leq 0$ , gives  $r^* \leq [r + K(s_0)] / [1 - K(s_0)]$ .

<sup>8</sup> Our banker will always be willing to extend some amount of credit if  $r^* > r^*$ . This is just an application of the standard expected utility diversification theorem which says that some part of a mathematically favorable gamble (investment) will always be taken when the choice is between a risky and a safe asset.

Differentiating (6) with respect to  $y$ , gives

$$\frac{\partial^2 L}{\partial y \partial z} = \int_{-1}^{\theta^*} V'' \{ \cdot \} (1 + \theta)(\theta - r) dF(\theta) \\ + V'(\cdot)(1 + \theta^*)f(\theta^*)y(1 + r^*)/(y + z)^2 > 0,$$

for  $V'' = 0$ . (Note that we would have  $\partial^2 L / \partial y \partial z > 0$ , for  $V'' \leq 0$ , provided that  $\theta^* < r$ , but for very high debt-equity ratios this latter condition could be violated.)

**PROPOSITION 3.** *Under constant returns, and with utility linear in wealth, the optimal loan is proportional to the borrower's equity, for  $0 < z^0 < w_0$ .*

Since maximization is on the interior the equality must hold in (6). For linear utility we set  $V' = 1$  in (6), and, integrating by parts, the supply function is defined implicitly by

$$G(\gamma, r^*, r) = r^* - r - (r^* - \theta^*)F(\theta^*) \\ (8) \quad - \int_{-1}^{\theta^*} F(\theta) d\theta = 0,$$

where

$$\theta^* = \frac{(1 + r^*)\gamma}{1 + \gamma} - 1, \quad \gamma = z^0/y$$

Since  $G$  depends only on  $\gamma$ , the ratio of optimal loan size to equity, and since  $G\gamma \neq 0$ , (8) defines a supply function of the form  $\gamma = (z^0/y) = s(r^*, r)$ . This explains why bankers may not just relate loan size to equity, but actually quote a fixed ratio at which they are willing to lend on the equity in an investment. (For example if bankers stand ready to make mortgage loans in the amount of two-thirds of the value of real estate, then  $\gamma = 2$ .) If bankers quote a fixed  $\gamma$  for a given class of loans (mortgages, consumer loans, or margin purchases) then this is evidence consistent with the hypotheses of linear utility functions and constant returns to scale.

#### IV. Characteristics of Lender Supply Function

The lender supply function defined by (6) or (8) may have the backward bending

characteristic of the J-M supply function. Thus, the supply function defined by (8) reaches a local maximum at  $r^* = \hat{r}^*$ , if

$$(d\gamma/dr^*) = [\gamma/\lambda(\theta^*)(\hat{r}^* - \theta^*)] \\ \cdot \{ [(1 + \gamma)/\gamma(\hat{r}^* - \theta^*)] - \lambda(\theta^*) \} = 0$$

or

$$(9) \quad \lambda(\theta^*) = (1 + \gamma)/\gamma(\hat{r}^* - \theta^*),$$

and

$$(d^2\gamma/dr^{*2}) = -[(\hat{r}^* - \theta^*)\lambda'(\theta^*)\gamma + \lambda(\theta^*)] \\ /[(\hat{r}^* - \theta^*)^2\lambda(\theta^*)]^2 < 0$$

or

$$(10) \quad (\hat{r}^* - \theta^*)\lambda'(\theta^*)\gamma + \lambda(\theta^*) > 0$$

In (10), the term  $\lambda(\theta^*) = f(\theta^*)/[1 - F(\theta^*)]$  is the conditional probability of default in the interval  $(\theta^*, \theta^* + d\theta^*)$ , or the *failure rate* as it is called in reliability theory.

We also have

$$(11) \quad (d\gamma/dr) = -(1 + \gamma)/(\hat{r}^* - \theta^*)f(\theta^*) < 0,$$

so that an increase in the opportunity cost of capital decreases the supply of risky debt finance.

It remains to verify that probability densities exist which satisfy conditions (8)–(10), i.e., produce backward bending supply. An easy example is the family of densities  $f(\theta) = \alpha e^{-\alpha(1+\theta)}$ ,  $\alpha > 0$ , for which  $F(\theta) = 1 - e^{-\alpha(1+\theta)}$  and  $\lambda(\theta) = \alpha$ . The supply function is defined by

$$(8') \quad (1 + r^*)/(1 + \gamma) = (1/\alpha) \\ + [1 + r - (1/\alpha)]e^{\alpha(1+r^*)\gamma/(1+\gamma)}$$

where  $r^* - \theta^* = (1 + r^*)/(1 + \gamma)$  in (8). From (9) we have

$$(9') \quad (1 + \hat{r}^*) = (1 + \gamma)^2/\gamma\alpha.$$

Substituting from (9') into (8') gives

$$1/[(1 + r)\alpha - 1] = \gamma e^{1+r}$$

which yields a finite  $\gamma = \hat{\gamma}$  for any  $\alpha > 1/(1 + r)$ .

Finally, since  $\lambda(\theta^*) = \alpha$ ,  $\lambda'(\theta^*) = 0$ , (10) is satisfied. Hence, for any  $\alpha > 1/(1 + r)$ , the supply function is a relative maximum at some finite  $\gamma = \hat{\gamma}(r)$  and,  $r^* = \hat{r}^*(r)$ .

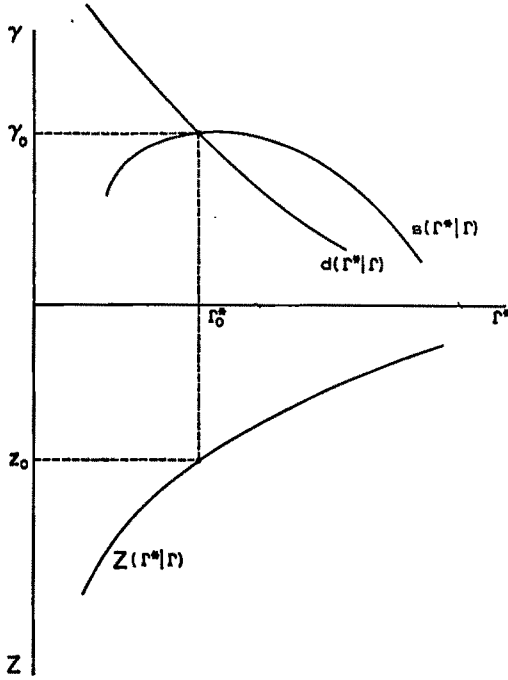


FIGURE 1

### V. The "Equilibrium" Loan, Pareto Optimality, and Capital Rationing

Now assume, as is common in ordinary markets, that the borrower and lender make their investment decisions independently at the market prices  $(r, r^*)$ . For given  $r$ , the lender's supply of debt per unit of equity is  $\gamma = s(r^*|r)$  as shown in the upper half of Figure 1. The borrower's demand for debt per unit of equity is  $d(r^*|r) = Z(r^*|r)/Y(r^*|r)$  from (3)–(4). This yields an equilibrium debt-equity ratio and interest rate  $(\gamma_0, r_0^*)$  as shown. At  $r_0^*$ , in the lower half of Figure 1, the loan size  $z_0$  is determined from the borrower's demand for debt,  $Z(r^*|r)$ .

But as is common in markets characterized by externality this market equilibrium is not in general mutually satisfactory to the two agents. Because the equity investment is an external economy to the borrower, he would like to see it increased, and would be willing to pay to see this accomplished. Any such change would then alter the borrower's decisions leading to a counter offer. Thus may further gains from trade be captured by each agent.

To develop these considerations more precisely, let  $B = B(u, z, y|r^*, r)$  be the borrower's criterion in (2), where  $u = (W_0 - y)$  is his investment in the safe asset. Similarly, let  $L = L(v, z|y, r^*)$  be the lender's criterion in (5), where  $v = w_0 - z$  is his investment in the safe asset. With independent action, each agent will maximize subject to his wealth constraint. For the borrower the marginal conditions are

$$(12) \quad B_u = B_v, \quad B_z = 0,$$

and the lender's condition

$$(13) \quad L_v = L_z$$

Now consider a Pareto optimal arrangement between the two agents. Their combined wealth resources are  $w_0 + W_0$ , to be allocated among  $u, v, y, z$ . In a closed general equilibrium system we might think of  $w_0 + W_0$  as society's initial endowment of seed corn to be divided between land with a sure yield of corn, and land with an uncertain yield of corn. The rules of the game are that investment  $z$  in the risky venture provides the lender with a priority claim on all yield up to a maximum of  $(1+r^*)z$ , while  $y$  gives the borrower all yield in excess of  $(1+r^*)z$ . The arrangement is constrained by the fact that the terminal distribution of wealth must follow this particular sharing plan. The Pareto criterion is to

$$\max_{u, v, y, z, r^*} B(u, z, y, r^*|r) + \xi L(v, z, y, r^*|r) + \mu(w_0 + W_0 - u - v - y - z)$$

giving the conditions

$$(14) \quad \begin{cases} B_u - \mu = 0, & B_v + \xi L_y - \mu = 0, \\ B_{r^*} + \xi L_{r^*} = 0, \\ \xi L_v - \mu = 0, & B_z + \xi L_z - \mu = 0 \end{cases}$$

It is clear that the "competitive equilibrium" of (12) and (13) which is illustrated in Figure 1 can never be Pareto optimal. Using (12) and (13), (14) can be satisfied only if  $L_v = 0$ , which violates Proposition 1.<sup>9</sup>

<sup>9</sup> This conclusion need not hold if the debtor's liability on the loan is not limited to the proceeds of the risky investment. Thus, suppose the proceeds of the

However, (14) is *entirely consistent with capital rationing*. To see this, suppose that our agents have hit upon a negotiated arrangement with the property that  $B_u > B_v$ ,  $B_z > 0$ , and  $L_v > L_z$ . Under such an arrangement,  $L_y = (B_u - B_v)(L_v/B_u) > 0$  and there is no contradiction. Furthermore, since  $B_z > 0$  (debt has a positive marginal utility), the borrower is rationed. These considerations suggest that free competitive negotiations between a borrower and lender may produce an arrangement which looks quite different from what is normally expected to characterize a "competitive equilibrium." But this difference is not a paradox once it becomes clear that the debtor-creditor contract is characterized by externality-like interdependence. This interdependence requires the negotiation process to include nonprice information in contrast with ordinary Pareto-Walras commodity markets.

The classical theoretical prescription for inducing Pareto optimality in markets with externalities is to internalize them with a system of charges and subsidies. An entirely equivalent device is to impose maximum or

debtor's investment in the safe asset  $(W_0 - y)(1 + r)$  are also available to the lender. Then the borrower's terminal wealth becomes zero, if

$$(1 + \theta)(y + z) + (W_0 - y)(1 + r) \leq (1 + r^*)z,$$

i.e., if  $1 + \theta \leq 1 + \theta^* = [(1 + r^*)z - (W_0 - y)(1 + r)] / (y + z)$ , while his terminal wealth is

$$(W_0 - y)(1 + r) + (1 + \theta)(y + z) - (1 + r^*)z,$$

if  $\theta > \theta^*$ . In place of (2), expected utility is now

$$B = \int_{\theta^*}^{\infty} U\{(W_0 - y)(1 + r) + (1 + \theta)(y + z) - (1 + r^*)z\} dG(\theta)$$

Since  $r^* > r$ , it follows that if  $B_y = 0$  for  $0 < y < W_0$ , then  $B_z < 0$ , and  $z = 0$ . An investor who desires to invest in the safe asset will not also want to borrow for risky investment. Similarly, if  $B_z = 0$ ,  $z > 0$ , then  $B_y > 0$ ,  $y = W_0$ . The investor will not wish to borrow until he has put all wealth in the risky asset. Therefore, in conditions (12)–(14) with positive borrowing, the borrower's expected utility depends only on the variable  $z$ ,  $B(0, z, W_0 | r^*, r)$ , and for the lender  $L = L(v, z | W_0, r^*)$ . Conditions (12) and (13) become  $B_z = 0$ , and  $L_v = L_z$ . For Pareto optimality, subject to the constraint  $0 \leq y \leq W_0$ , conditions (14) become  $B_u - \mu < 0$  ( $\mu = 0$ ),  $B_y + \xi L_v - \mu > 0$  ( $y = W_0$ ),  $B_{r^*} + \xi L_{r^*} = 0$ ,  $\xi L_v - \mu = 0$ ,  $B_z + \xi L_z - \mu = 0$  which are satisfied by the modified conditions (12) and (13).

minimum quotas on the appropriate decision variables. In fact, a charge system has a dual relationship to a quota system in which the shadow value of the "right" quota is precisely the right tax. A Pareto optimal smoke "rationing" quota has a shadow price equal to the Pareto optimal smoke tax. This suggests the possibility that capital rationing is a negotiated, decentralized, quota system. The conditions (12)–(13) for a "competitive equilibrium" clearly do not exhaust the gains from trade, while a quota system, if it satisfies (14), can exhaust such gains.

An example of a quota or rationing system that would be Pareto optimal is the following. Let the lender (who always prefers an increase in equity) set a minimum quota  $\hat{y} \leq y$  on the borrower's equity in the project, and let the borrower set a minimum quota  $\hat{z} \leq z$  on the debt capital supplied by the lender. Now suppose the institutional arrangement is one in which the borrower having specified  $\hat{z}$ , chooses  $(u, y)$  subject to the lender's quota,  $\hat{y}$ . Similarly, the lender having specified  $\hat{y}$ , chooses  $(v, z)$  subject to the borrower's quota,  $\hat{z}$ . The borrower will

$$(15) \quad \max_{u, y} B(u, y | \hat{z}, r^*, r) + \alpha(w_0 - y - u) + \alpha'(y - \hat{y})$$

giving

$$(16) \quad B_y - \alpha + \alpha' = 0, \quad B_u - \alpha = 0$$

The lender will

$$(17) \quad \max_{v, z} L(v, z | \hat{y}, r^*, r) + \beta(w_0 - z - v) + \beta'(z - \hat{z})$$

giving

$$(18) \quad L_v - \beta = 0, \quad L_z - \beta + \beta' = 0$$

Conditions (16) and (18) correspond to Pareto optimality provided that the quotas  $(\hat{y}, \hat{z})$  are such that the shadow prices  $(\alpha', \beta')$  have the values  $\alpha' = \xi L_y$ ,  $\beta' = B_z / \xi$ . Thus, in equilibrium, the borrower's quota restraint on equity must have a shadow value which reflects the lender's marginal utility of equity, while the lender's loan quota must

have a shadow value which reflects the borrower's marginal utility of debt finance.<sup>10</sup>

This development by no means deserves to be called a negotiation process, for it begs the negotiation question of how  $y$ ,  $z$  are determined, just as the classical charge-subsidy scheme begs the question of how to determine the optimal levies. What is clear is that capital rationing can be optimal in the loan market where uncertainty gives rise to

<sup>10</sup> The relation between these shadow values and externality charges can be seen if we imagine that the lender makes a payment  $e$  per dollar of equity which is transferred to the borrower. The borrower pays a unit tax  $d$  on debt capital that is transferred to the lender. The borrower must now max  $B(u, y|z, r^*, r) + \alpha[w_0 - (1-e)y - dz - u]$  giving conditions (16), if  $\alpha' = \alpha e$ . The lender will

$$\max L(v, z|y, r^*, r) + \beta[w_0 - ey - (1-d)z - v]$$

giving conditions (18), if  $\beta' = \beta d$ .

priority-sharing legal arrangements, and makes every loan a unique contract. Of course, transactions and information cost may force lenders to group loans into fairly broad risk classes within which lending rules may be fairly uniform.

# REFERENCES

- D. M. Jaffee, and F. Modigliani, "A Theory and Test of Credit Rationing," *Amer. Econ. Rev.*, Dec. 1969, 59, 850-72.
- J. P. Quirk, "The Capital Structure of Firms and the Risk of Failure," *Int. Econ. Rev. Proc.*, May 1961, 51, 210-28.
- V. L. Smith, "Investment Decision, Uncertainty and the Incorporated Entrepreneur," in J. Quirk and A. Zarley, eds., *Papers in Quantitative Economics*, II, Kansas City 1971.

# A Theory and Test of Credit Rationing: Further Notes

By DWIGHT M. JAFFEE\*

Vernon Smith's paper provides an interesting and important generalization of the theory of credit rationing developed in Jaffee and Modigliani (J-M). The theory developed in our paper was based, as a simplifying assumption, on the premise that the borrowing firm's end of period value was independent of the size of the loan granted by the bank. Smith's paper generalizes our results to include the case in which the borrower's investment and hence his end of period value is explicitly a function of the size of the loan granted by the bank. For purposes of reference, the former case will be called the *fixed-size assumption* and the latter case the *variable-size assumption*.

In this note, I will elaborate on two issues raised by Smith. First, although Smith develops the implications of the variable-size assumption for the shape of the bank's optimal loan offer curve, he only briefly considers the extent to which this assumption modifies the basic results of J-M. It will be shown in Section I below that, in fact, the main properties of the offer curve are left completely unchanged. Second, Smith develops algebraically the consistency of credit rationing with Pareto optimal loan contracts. In Section II below, a diagrammatic approach is used to illustrate the relationship between credit rationing and Pareto optimal loan contracts, and then the positive aspects of actual loan negotiations are contrasted with the normative conditions of the Pareto optimal frontier.

It should perhaps be added, before turning to these issues, that Smith's contribution takes on further significance in view of the literature that has been concerned with the variable-size assumption. The assumption of a fixed-size investment opportunity was first introduced in the path-breaking study of

credit rationing by Donald Hodgman (1960). The assumption was then critically discussed in the exchange between Sam Chase and Hodgman (1961), and a preference for the variable-size assumption was acknowledged. The analytic properties of the bank offer curve under the fixed-size and variable-size assumptions were next developed further by Marshall Freimer and Myron Gordon. Their analysis was based, however, on the special case where the distribution of possible outcomes is given by a rectangular density function, and was marred by an untenable assumption concerning the relationship of the bank's opportunity cost and the expected value of the borrower's investment project.<sup>1</sup> Finally, in J-M, we simplified the exposition by using the fixed-size assumption, but noted that the more general case for a variable-size assumption was developed in Jaffee (1968).<sup>2</sup> As it turns out, the general approach adopted by Smith parallels my analysis of the problem.

## I. The Bank Offer Curve with the Variable-Size Assumption

In order for the analysis of credit rationing developed in J-M to remain valid under the variable-size assumption, the essential condition is that the three main characteristics of the offer curve remain unchanged. These characteristics may be summarized:

- (i) the optimal loan offer is zero at some nonzero loan rate and then rises with the loan rate over some range;
- (ii) the optimal loan offer reaches a finite global maximum;
- (iii) as the interest rate approaches infinity,

<sup>1</sup> The nature of this assumption and its relationship to the Freimer and Gordon analysis is given in fn. 4 below.

<sup>2</sup> See Jaffee and Modigliani, p. 852 and fnn. 5 and 11. The discussion in Jaffee (1968) has been revised and expanded in Jaffee (1971).

\* Department of economics, Princeton University.

the loan offer declines and, in special cases, approaches zero.

Smith has developed the validity of property (i), but he does not consider properties (ii) and (iii).

A proof of property (ii), using Smith's notation but our own analysis, can be developed as follows. Equation (6) in Smith's paper can be reinterpreted as giving the first-order condition for the maximization of expected profits ( $L$ ) with respect to loan size ( $z$ ), and can be written:

$$(1) \quad \frac{\partial L}{\partial z} = (r^* - r) - (1 + r^*)F[\theta^*] + \int_{-1}^{\theta^*} (1 + \theta)f(\theta)d\theta,$$

where

$$\theta^* = [\gamma(1 + r^*)/(1 + \gamma)] - 1, \quad \gamma = z/y,$$

and where  $V'[\cdot] = 1$  and  $f[-1] = 0$  to be consistent with J-M.<sup>3</sup> The profit maximum is obtained when the marginal profit function (1) is set equal to zero. To show that this maximum cannot occur at an infinite loan size, it must be proven that:

$$(2) \quad \lim_{z \rightarrow \infty} \frac{\partial L}{\partial z} < 0$$

If this condition is valid, then the bank would find its profits decreasing as the loan offer approaches infinity.

Now, from the definition in equation (1) we have:

$$(3) \quad \lim_{z \rightarrow \infty} \theta^* = r^*,$$

and arranging (1) and using (3), we have:

$$(4) \quad \lim_{z \rightarrow \infty} \frac{\partial L}{\partial z} = (1 + r^*)(1 - F[r^*]) + \int_{-1}^{r^*} (1 + \theta)f[\theta]d\theta - (1 + r)$$

<sup>3</sup> J-M assumes the bank is a simple expected profit maximizer and that some minimum nonzero return is assured on the firm's investment.

Condition (2) ruling out infinite loans can then be written, on the basis of equation (4), as:

$$(5) \quad (1 + r^*)(1 - F[r^*]) + \int_{-1}^{r^*} (1 + \theta)f[\theta]d\theta < (1 + r)$$

Condition (5) means that, for a given loan rate  $r^*$ , the expected value of the loan contract to the bank must be less than the bank's opportunity cost  $(1+r)$ . An interpretation of this condition follows directly from the definition of the opportunity cost. For example, assume condition (5) does not hold. The bank could then invest its entire portfolio in loans to this customer and still obtain an expected return exceeding the opportunity cost. Thus, contrary to assumption,  $(1+r)$  would not be the opportunity cost; instead, the bank would use the expected value of this loan as the standard for judging the profitability of other investment alternatives.<sup>4</sup>

Given that the opportunity rate  $(1+r)$  is finite, it is clear that condition (5) will necessarily hold only if further conditions are placed on the expected value of the loan contract. In my earlier analysis, two conditions on the properties of the borrower's outcome function were suggested. First, it could be assumed that the maximum possible outcome for the borrower's firm is finite; that is, there exists a finite number  $K$  such that:

$$(6) \quad f[\theta] = 0 \quad \text{for } \theta \geq K$$

This insures that the expected value of the loan contract will be finite. Alternatively, it could be assumed that the expected rate of profit on the borrower's investment project exhibits diminishing returns as the scale of the project (that is, the bank's investment) increases.

<sup>4</sup> In the analysis of Freimer and Gordon, referred to above, the shape of the offer curve was derived by simple numerical analysis for the case of a rectangular density function. In all their numerical examples, however, the expected value of the project exceeded the bank's opportunity cost, and thus they concluded that the offer curve would become infinite at some point.

Turning now to property (iii) of the loan offer curve, assume that conditions (5) and (6) are valid. The derivative of the profit function with respect to loan size as the loan rate approaches infinity can be derived, using condition (6), from equation (4):

$$(7) \quad \lim_{r^* \rightarrow \infty} \frac{\partial L}{\partial z} = \int_{-1}^K \theta f[\theta] d\theta - r$$

This function is independent of the loan size and, thus, if it is positive, the optimal loan is infinite, and, if it is negative, the optimal loan is zero. Condition (5), of course, insures that the function is negative, and thus the optimal loan approaches zero as the loan rate approaches infinity.

Intuitively, the result is quite understandable. As the loan rate approaches infinity, the bank is able to claim the entire proceeds of the borrower's investment project, including that portion based on the borrower's equity. The bank could obtain even further revenue if it increased the scale of the project above the minimum provided by the borrower's equity, but the marginal profits to the bank, given by equation (7), would then be negative. Consequently, in the limit the bank will settle for the proceeds based on the borrower's equity, and its own contribution in the form of loan funds will approach zero. It can also be noted that a somewhat more general result is obtained if the assumption of condition (6) is replaced by the assumption of a diminishing return to scale for the expected rate of profit on the borrower's investment project. In this case, instead of the zero-infinity choice for the loan size when the interest rate approaches infinity, a continuous solution is obtained in which the bank invests in the project up to the point at which the expected profit rate of the project just equals the bank's opportunity rate.

## II. Credit Rationing, Pareto Optimality, and Actual Loan Negotiations

In the concluding section of his paper, Smith provides an analysis of credit rationing under conditions of Pareto optimality in the loan market. The analysis is restricted, however, in that the relative degree of market power attributed to the banks is not

made explicit, and, consequently, the implications of varying degrees of bank market power cannot be developed. In this section the Smith analysis is extended to allow a diagrammatic consideration of these questions.

As a starting point, it is useful to extend the demand curve-offer curve diagram, shown as Figure 1 in Smith's paper, to incorporate the iso-profit mappings of the bank and of the borrower. This diagram is shown in my Figure 1. The offer curve of the bank,  $S$ , and the demand curve of the borrower,  $Z$ , drawn with darker lines in the figure, represent the optimal loan quantities at each interest rate. The iso-profit curves for the bank are shown as the family of curves,  $S_0, S_1, S_2, S_3$ , where the subscripts indicate the relative ordering in terms of expected profits, and  $S_0$  is the zero profit locus. The characteristics of these iso-profit curves are: (i) they are vertical where they intersect the offer curve; (ii) they are positively sloped where they are above the offer curve; and (iii) they are negatively sloped where they are below the offer curve. The vertical slope of the iso-profit curves follows directly from the definition of the offer curve as the locus of optimal loan sizes for varying interest rate factors.<sup>5</sup> The slope (and concavity) of the iso-profit curves is based on Proposition 2 of J-M, namely that expected profits decrease monotonically as the loan size varies from the optimal size in either direction. The iso-profit curves for the borrowing firm are shown as  $Z_0, Z_1, Z_2, Z_3$ , where, again, the subscripts indicate the relative ordering in terms of profits, and  $Z_0$  is the zero profit locus. The shape of the borrower's iso-profit curves and their relationship to the borrower's demand curve are assumed to be symmetrical to the corresponding curves for the bank.<sup>6</sup>

A number of possible solutions to the loan contract negotiation can be illustrated with Figure 1. To start, the Pareto optimal con-

<sup>5</sup> If the iso-profit curves were not vertical at the intersection with the offer curve, then a vertical move away from the offer curve would lead to higher profits, contrary to the definition of the offer curve.

<sup>6</sup> A more rigorous algebraic derivation of the properties of the bank's iso-profit curves is provided in Jaffee (1971), Section (2.3.3).

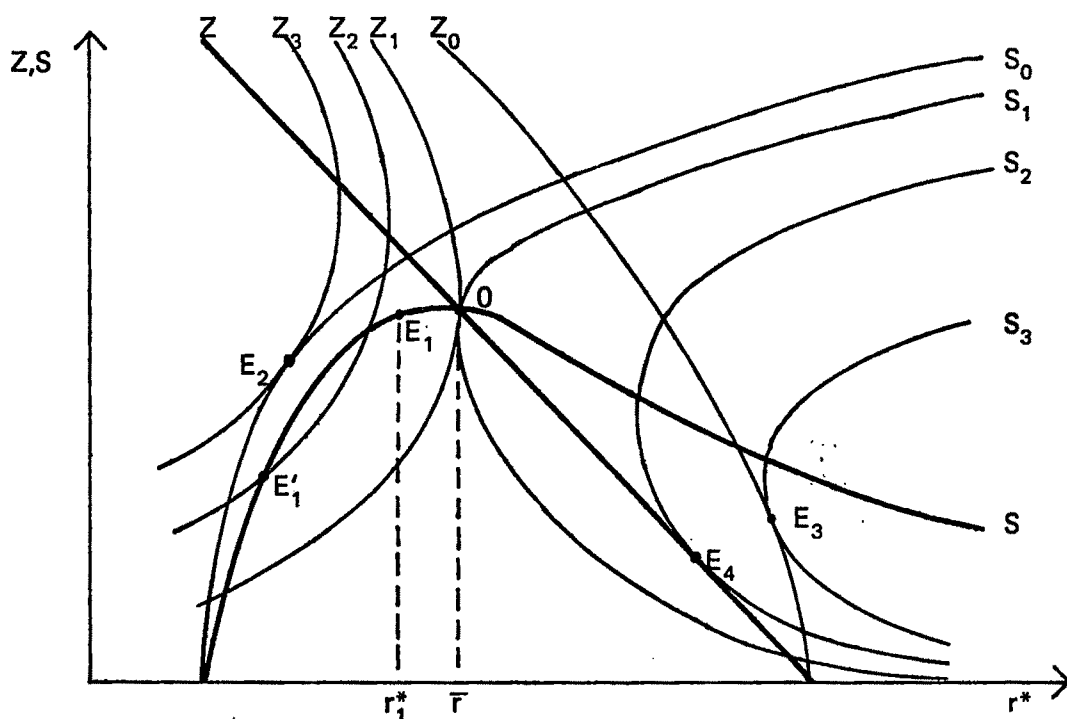


FIGURE 1

tracts will consist of those points at which the iso-profit curves of the bank and the borrower are tangent. At such points the expected profits of one of the parties will be maximized given the level of expected profits for the other party. Three of these points are specifically illustrated in the diagram: at  $E_2$  the borrower's expected profits are constrained only by the zero profit locus of the bank; at  $E_3$  the bank's expected profits are constrained only by the zero profit locus of the borrower; and at  $0$ , which is a more "neutral" point, both iso-profit curves are vertical and thus intersect the respective offer and demand curves.

Two properties of the Pareto optimal locus are worth noting:

**PROPERTY 1.** *The Pareto optimal locus must always lie between the bank's offer curve and the borrower's demand curve; furthermore, to the left of the intersection of the demand curve and the offer curve (shown at the rate  $\bar{r}$ ) the Pareto optimal locus is below the demand curve, and to the right of this intersection the Pareto optimal locus is below the offer curve.*

This property of the Pareto optimal locus follows directly from the shape of the respective iso-profit mappings. The property is important because it illustrates the consistency of credit rationing with Pareto optimal loan contracts, as noted by Smith. Specifically, for all interest rates below  $\bar{r}$  (in the figure), the Pareto optimal loan is less than the loan demanded by the borrower. It should be stressed, however, that for the remainder of the Pareto locus, where the interest rate is above  $\bar{r}$ , it is the bank, not the borrower, who is rationed.<sup>7</sup> Thus the restriction that the quoted interest rate lie below  $\bar{r}$  is as necessary for the credit rationing of the borrowers of Pareto optimal loans as it is for the rationing of borrowers under the various market conditions described in J-M.<sup>8</sup>

<sup>7</sup> Thus it is only at the rate  $\bar{r}$  that both the borrower and the bank are satisfied with the Pareto optimal loan.

<sup>8</sup> For example, a banker acting as a discriminating monopolist and maximizing its expected profits with respect to the borrower's demand function would choose point  $E_4$  in the figure for its loan contract. The interest rate on this contract is necessarily above  $\bar{r}$ , and

PROPERTY 2. *The bank's expected profits rise and the borrower's expected profits decrease along the Pareto optimal locus as the interest rate rises.*

This property of the locus is based directly on the ordering of the iso-profit mappings. A consequence of this property is that banks will prefer Pareto optimal loans with high interest rates and that borrowers will prefer Pareto optimal loans with low interest rates. While this result may appear obvious, it does help to stress, along with Property 1, that the existence of credit rationing, even with Pareto optimal loans, still depends on the relative market power of the bank. In particular, if, as commonly assumed, banks do possess the major power in the loan market, then they would always prefer to choose loans with very high interest rates and credit rationing would not be expected to occur.

It thus follows that additional features of the loan market must be introduced if credit rationing is to be a likely outcome of loan negotiations. One set of such features of the loan market are the legal and social constraints that make it difficult for banks to discriminate blatantly between loan customers. It is apparent that solutions near the  $E_3$  end of the Pareto optimal locus would, in fact, entail such discrimination. For this reason, one would expect the banks to exert their market power in more subtle ways. The details of one system of customer classification, with objective criteria for the definition of classes that could be used for this purpose, have already been discussed in J-M. For the purposes of the present discussion, it need only be noted that the result of such a classification system is that customers are

quoted specific loan rates, and that the main issue of negotiation between the bank and the borrower is the size of the loan to be granted. This can be illustrated in Figure 1, where it is assumed that the loan rate for the particular borrower's class is  $r_1^*$ , which lies below  $\bar{r}$ .<sup>9</sup> The bank would then offer the loan size shown as  $E_1$ , and since this is less than the loan demand at  $r_1^*$ , the customer would be rationed. In practice, of course, negotiations could lead to a loan size somewhat above  $E_1$ , possibly even the Pareto optimal loan size, but credit rationing would still occur as long as the contracted amount lies below the demand function.

#### REFERENCES

- S. B. Chase, Jr., "Credit Risk and Credit Rationing: Comment," *Quart. J. Econ.*, May 1961, 75, 319-27.
- M. Freimer and M. Gordon, "Why Bankers Ration Credit," *Quart. J. Econ.*, Aug. 1965, 79, 397-416.
- D. R. Hodgman, "Credit Risk and Credit Rationing," *Quart. J. Econ.*, May 1960, 74, 258-78.
- , "Credit Risk and Credit Rationing: Reply," *Quart. J. Econ.* May 1961, 75, 327-329.
- D. M. Jaffee, "Credit Rationing and the Commercial Loan Market," doctoral dissertation, M.I.T. 1968.
- , *Credit Rationing and the Commercial Loan Market*, New York 1971.
- and F. Modigliani, "A Theory and Test of Credit Rationing," *Amer. Econ. Rev.*, Dec. 1969, 59, 850-72.
- V. Smith, "The Theory of Credit Rationing—Some Generalizations," *Amer. Econ. Rev.*, June 1971, 61, 477-83.

thus the banker would not ration credit even if given the opportunity to do so; indeed, as in standard theory, the discriminating monopolist always would prefer to make a larger loan at the monopolist interest rate.

<sup>9</sup> As discussed in J-M, cases in which the quoted rate falls below the rate  $\bar{r}$  are likely to occur under such a classification scheme.

# Distributional Equality and Aggregate Utility: Further Comment

By MAURICE McMANUS, GARY M. WALTON, AND RICHARD B. COFFMAN\*

The recent communication in this *Review* by William Breit and William Culbertson raises several intriguing and important questions about Abba Lerner's theorem regarding the optimum distribution of income. Most previous criticisms of Lerner's theorem, such as those by I. M. D. Little, pp. 57-66, Milton Friedman, pp. 308-10, and Richard Musgrave, pp. 106-09, focused on the validity of its assumptions.<sup>1</sup> Breit and Culbertson's contribution is novel in that it is the first criticism directed explicitly at the logical consistency and completeness of Lerner's proof.

In attacking Lerner's proof, Breit and Culbertson first distinguish between two versions of the theorem: "One of them is the relatively meek and mild assertion that '... the maximization of *probable* total satisfaction is attained by an equal division of income' (p. 29). The other is the much bolder claim '*... that if it is desired to maximize the total satisfaction in a society, the rational procedure is to divide income on an equalitarian basis*' (p. 32, Lerner's italics)," (Breit and Culbertson, p. 438).

After reviewing Lerner's proof, Breit and Culbertson conclude that "Lerner has proved his weak proposition for a two-person model while asserting his bolder one for an *N*-person society" (p. 439). In Section III of their communication, Breit and Culbertson: 1) offer proofs that Lerner's strong proposition does not hold in either a two-person or an *n*-person world, and 2), assert in a footnote that even Lerner's weak proposition fails for an *n*-person world unless it is assumed "... that the same degree of inequality

holds for all individuals throughout society" (p. 439, fn. 8).

At that point Breit and Culbertson have left Lerner with only a proof of his weak proposition as applied to a two-person world. However, they immediately declare themselves to be in agreement with the conclusion of Lerner's strong theorem for an *n*-person society, and proceed in Section IV to present a proof of that theorem.

In Lerner's reply to Breit and Culbertson, he denies having intended to put forth two propositions, and offers clarification of the single sentence which led Breit and Culbertson to their discovery of the bold proposition.<sup>2</sup> More importantly, he disavows any belief in the bold proposition, and he argues that Breit and Culbertson's proof is neither necessary for the weak proposition, nor sufficient for the strong one.

Early in their communication Breit and Culbertson correctly point out that no rigorous proof of Lerner's theorem exists.

<sup>2</sup> Lerner states: "Perhaps the ambiguity would have been avoided if I had added the following words in Roman type to the italicized sentence quoted: '*... if it is desired to maximize the total satisfaction in a society, the rational procedure*, in the absence of the knowledge that would enable us to equalize the marginal utilities, *is to maximize the probable total satisfaction—i.e., to divide income on an equalitarian basis*.'" (1970, p. 442).

Evidently Breit and Culbertson also find the bold proposition in Lerner's argument about a large population (see p. 492 below). After quoting Lerner, Breit and Culbertson give what is apparently their interpretation of his conclusion: "... in the aggregate, however, the losses would be greater than the gains," and they promise to show that "even in this aggregate case, he cannot assert this bold claim" p. 439.

We, on the other hand do not interpret Lerner's argument as a claim that a welfare gain *would* result from redistribution in a large society. We find only an informal, Law-of-Large-Numbers argument that when many rather than few pair-wise redistributions are made, the proportion of "successful" redistributions (as observed, perhaps, by God) is less likely to depart from the expected 50 percent.

\* Professor of economics, University of Birmingham, England, associate professor of economics, Indiana University, and acting assistant professor of economics, University of Hawaii, respectively.

<sup>1</sup> Lerner's five assumptions are lucidly stated by Breit and Culbertson, pp. 435-36.

The exchange between Lerner and Breit and Culbertson does not appear to have remedied the situation. Lerner's rejection of their proof and their rejection of his leaves a void which must be filled if Lerner's theorem is to stand on a logical foundation. The primary purpose of this note is to provide that foundation.

In Section I we offer a rigorous proof of Lerner's weak theorem for the  $n$ -person case. We then contrast this with Breit and Culbertson's propositions and clarify and disprove their rejection of his theorem. Section I is concerned entirely with the logical consistency of Lerner's theorem, given his assumptions. Our emphasis there on the issue of logical consistency to the exclusion of all else should not be interpreted as an unquestioning acceptance of Lerner's assumptions. To the contrary, we share the uneasiness which previous critics have expressed. In Section II we address ourselves to the equal ignorance assumption, showing that Lerner's egalitarian conclusion can be

derived from some alternative assumptions about ignorance.

### I. Proof of Lerner's Weak Theorem in the $N$ -Person Case

Lerner's (only) theorem in the two-person case can be summarized as follows. Figure 1 is similar to that used by Lerner and by Breit and Culbertson. Starting from equal incomes of \$100, there are two possible outcomes for a redistribution to, say, \$120 and \$80. If the person with \$120 is also the person with marginal utility schedule  $A$  then there will be a net gain of utility of area  $G$  since he gains  $F+G$  and the other person loses  $F$ . On the other hand, if schedule  $A$  belongs to the poor person then the redistribution leads to a net loss of  $L$  since he loses  $L+K$  and the gainer gains  $K$ . In the absence of any knowledge of how utility functions are allocated between the individuals, each of the possible outcomes is considered equally likely. Under this 'Bayesian' assumption the

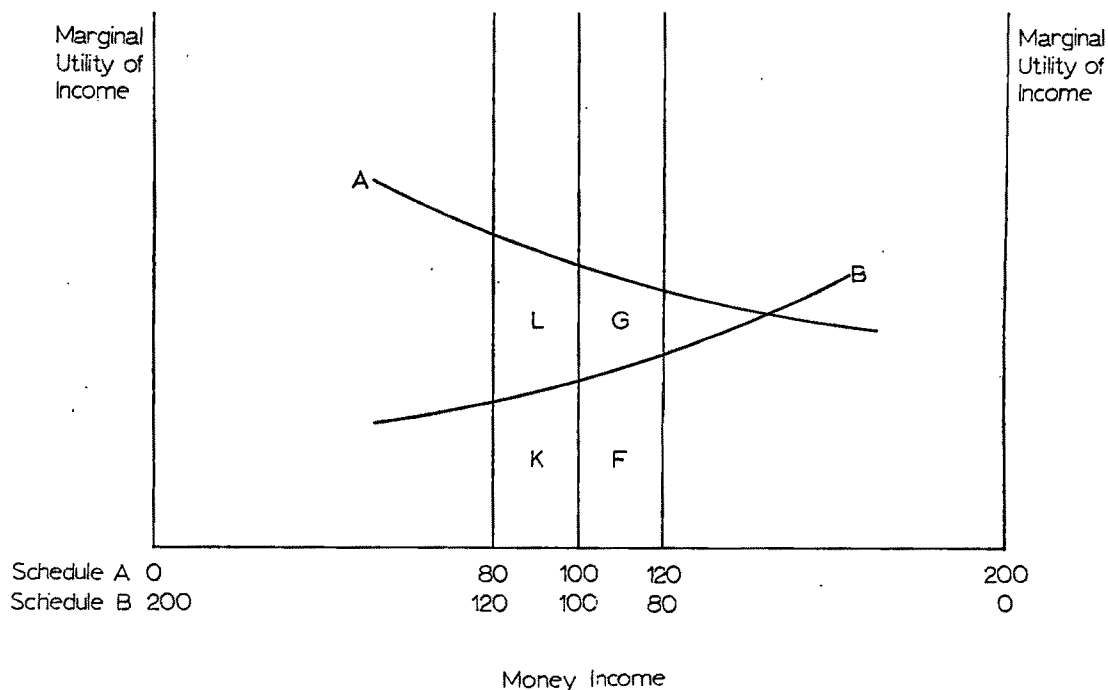


FIGURE 1

expected welfare change is  $(1/2)G - (1/2)L$  which is negative since  $L > G$ . Since the same argument applies to any redistribution, egalitarianism is proved to be optimal in Lerner's sense.

Lerner, p. 32, asserts that his theorem extends to the  $n$ -person case by arguing that gains like  $L$  and losses like  $G$  will occur with approximately equal frequency when the number of redistributions to equality is large. Similarly, Paul Samuelson, p. 175, gives a neat algebraic proof of the two-person case and asserts that the conclusion "holds for any number of summed concave (utility) functions." Unfortunately, the subsequent criticisms by Breit and Culbertson indicate that the generalization is not as self-evident as Samuelson suggests, and so an explicit formal proof of an  $n$ -person theorem is now offered.<sup>3</sup>

Consider a society of  $n$  members, with mean income  $\bar{Y}$ . If income inequality exists, individual incomes may be written as

$$(1) \quad \begin{aligned} & (a) \quad (\bar{Y} + h_1), \dots, (\bar{Y} + h_n) \\ & (b) \quad \sum_{j=1}^n h_j = 0, \end{aligned}$$

where the  $h_j$  are the individual deviations from the mean. Let their (cardinal) utility functions be denoted by

$$(2) \quad F_i(Y), \quad i = 1, \dots, n$$

We have insufficient knowledge to know which person has which utility function, so we make the Bayesian assumption that each possibility is equally likely. Given this assumption the mathematical expectations of the utilities of each income are

<sup>3</sup> Apparently Breit and Culbertson believe that the Lerner and Samuelson proofs do not generalize to the  $n$ -person case because unequal distribution between two persons involves only deviations from the mean which are equal though opposite in sign, whereas more complicated patterns are possible with  $n > 2$  persons. Our proof, which shows this reasoning to be erroneous, is designed to be closely analogous to the simple case in the sense that we average the negative and positive deviations separately and then note that the two averages have the same properties as incomes in the two-person case.

$$(3) \quad \frac{1}{n} \sum_i F_i(\bar{Y} + h_j), \quad j = 1, \dots, n$$

Social welfare is considered to be the sum of the individuals' utilities. The mathematical expectation of social welfare is thus

$$(4) \quad \frac{1}{n} \sum_i \sum_j F_i(\bar{Y} + h_j)$$

This double summation may also be arrived at by a slightly different route. There is an equal chance of a given utility function being associated with each of the possible incomes, so that the expected utilities for the separate functions are

$$(5) \quad \frac{1}{n} \sum_j F_i(\bar{Y} + h_j) \quad i = 1, \dots, n$$

Of course, these sum to (4) again. This latter route to (4) is the one used in the argument which follows.

Let the  $h_j$  be numbered so that

$$(6) \quad h_1 \leq \dots \leq h_m \leq 0 \leq h_{m+1} \leq \dots \leq h_n$$

Evaluate the slope of  $F_i$  between the mean income value and each of the possible income levels. Since marginal utility is assumed to be diminishing, the function is concave and so the gradient decreases as the value of the argument increases. Hence

$$(7) \quad \begin{aligned} & \frac{F_i(\bar{Y}) - F_i(\bar{Y} + h_1)}{-h_1} \geq \dots \\ & \geq \frac{F_i(\bar{Y}) - F_i(\bar{Y} + h_m)}{-h_m} \\ & \geq \frac{F_i(\bar{Y} + h_{m+1}) - F_i(\bar{Y})}{h_{m+1}} \geq \dots \\ & \geq \frac{F_i(\bar{Y} + h_n) - F_i(\bar{Y})}{h_n} \end{aligned}$$

Next compute the weighted average of the first  $m$  slopes in (7), using weights proportional to the absolute values of the corresponding  $h_j$ , and also calculate the correspondingly weighted average of the remain-

ing slopes. The first average cannot be smaller than the smallest of the relevant slopes, i.e., the  $m$ th one, and similarly, the second average is at least as small as the  $(m+1)$ th slope. Hence (7) shows that

$$(8) \quad - \sum_{j=1}^m [F_i(\bar{Y}) - F_i(\bar{Y} + h_j)] / \sum_1^m h_j \\ \geq \sum_{j=m+1}^n [F_i(\bar{Y} + h_j) - F_i(\bar{Y})] / \sum_{m+1}^n h_j$$

The denominators in (8) are positive and equal in view of (6) and (1b). Cancelling them out and collecting like terms,

$$(9) \quad nF_i(\bar{Y}) \geq \sum_{j=1}^n F_i(\bar{Y} + h_j)$$

The same argument applies to each function. Dividing by  $n$  and summing the results over  $i$ ,

$$(10) \quad \sum_i F_i(\bar{Y}) \geq \frac{1}{n} \sum_i \sum_j F_i(\bar{Y} + h_j),$$

which states precisely that the expected welfare of the group is, if anything, larger for an equal income distribution.<sup>4</sup> The result is given for weakly concave functions. If, in addition, at least one of the functions is strictly concave then the result is sharpened to the strict inequality for not all  $h_j$  equal to zero.

Figure 2 illustrates the three-person case. It is now more convenient to measure all incomes in the same direction, thus marginal utility schedules  $A$ ,  $B$ , and  $C$  all have negative slopes.<sup>5</sup> Incomes are \$50, \$80, and \$170. It is assumed that each individual's utility function is equally likely to be associated with each income. The \$70 positive deviation from mean income is divided at the \$150 income level to facilitate comparison with the \$20 and \$50 negative deviations associated with the other two incomes. If incomes are redistributed to equality:

<sup>4</sup> The result only depends on concavity of the functions and holds even if marginal utilities can become negative.

<sup>5</sup> If the curves cut, the areas are more complicated but the results are the same.

$$\Delta E(W) = \left(\frac{1}{3}\right) \cdot \left\{ (G_1 + H_1) + H_1 - (J_1 + K_1) \right. \\ (11) \quad \left. + \sum^2 (G_i + H_i) + \sum^2 H_i - \sum^2 (J_i + K_i) \right. \\ \left. + \sum^3 (G_i + H_i) + \sum^3 H_i - \sum^3 (J_i + K_i) \right\}$$

Within the brackets the possible effects of redistribution are arranged so that line  $j$  refers to the  $j$ th utility function in the following fashion: line 1 shows that if the person with \$50 has marginal utility schedule  $A$ , redistribution will give a gain of area  $G_1 + H_1$ ; if the person with \$80 has utility schedule  $A$  then redistribution will give a gain of area  $H_1$ ; if  $A$  belongs to the person with \$170, redistribution will reduce welfare by  $J_1 + K_1$ . There is, under the Bayesian assumption, a one-third chance of each possibility. It is visually apparent that each  $\sum (G_i + H_i)$  area is greater than the corresponding  $\sum J_i$  area and each  $\sum H_i$  area is greater than the corresponding  $\sum K_i$  area. It follows immediately from (11) that income equality increases expected welfare.

This explicit proof of the theorem in the  $n$ -person case makes it easier to appraise the statements by Breit and Culbertson. They agree that the two-person proof will generalize to an  $n$ -person world if "Lerner explicitly assumes that the same degree of inequality holds for individuals throughout society" (p. 439), but not if incomes are unequally distributed around the mean.

Since Breit and Culbertson's view is very tersely stated we are hesitant to interpret their argument. However, judging by its context, we suspect it is based on logic similar to that employed in their rejection of the strong position:

Consider 100 million islands . . . each containing the same number of coconuts and inhabited by two men,  $A$  and  $B$ , with direct communications to Lerner. If Lerner were to receive a call from each asking him how to divide the coconuts, he would say, "Divide evenly." But this advice will certainly not do in the aggregate. For now he has no way of knowing that in the 50 percent of

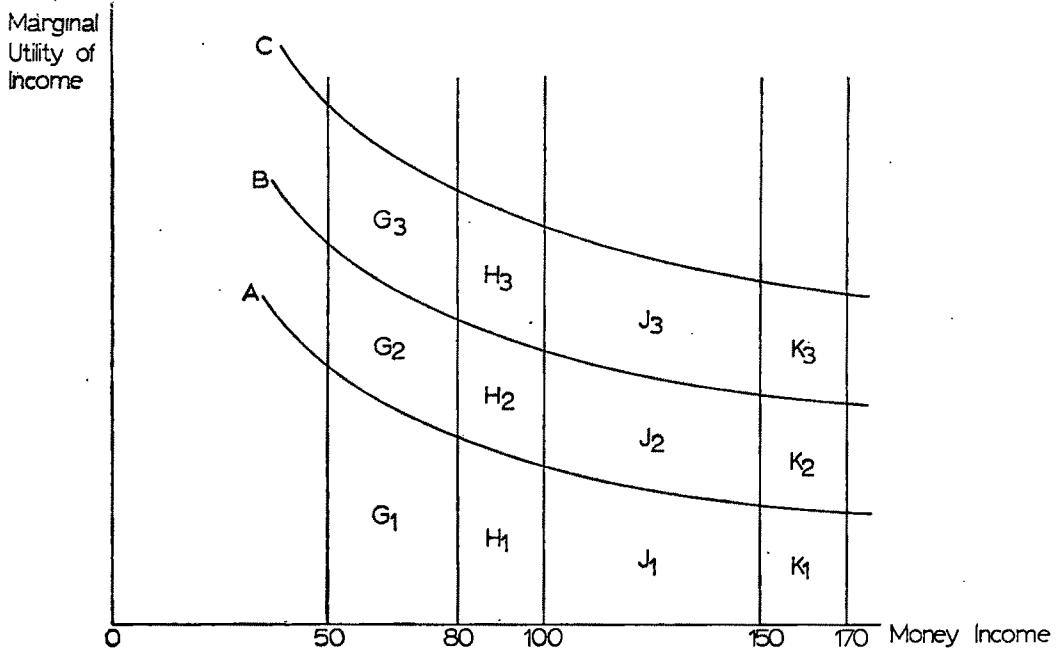


FIGURE 2

those instances when gains were made, they were greater than the losses in those 50 percent of the instances when losses were made. For without any prior knowledge about the utility functions of the 200 million individuals he cannot compare the areas under the marginal utility curves of the different individuals on the different islands. [p. 439]

This quotation makes clear one fundamental reason why Breit's and Culbertson's conclusions are at variance with the theorem just proved: they are examining a completely different proposition. Although (4) would seem to be the most "natural" generalization of the Lerner-Samuelson concept of expected welfare, Breit and Culbertson interpret this generalization to mean the case in which  $n$  is an even number, the population is divided into  $n/2$  pairs, all having identical mean incomes, and in which a pair of utility functions is associated with each pair of persons, but it is not known which person of each pair has which function. Instead of (4), expected social welfare would therefore be

$$(a) \frac{1}{2} \left[ \sum_{i=1}^2 \sum_{j=1}^2 F_i(\bar{Y} + h_j) \dots + \sum_{n-1}^n \sum_{n-1}^n F_i(\bar{Y} + h_j) \right]$$

$$(12) \quad (b) h_1 + h_2 = \dots = h_{n-1} + h_n = 0$$

Since neither one of (4) and (12) implies the other, there is no reason to expect conclusions based on them to be the same. Even so, it is easy to see that, even within their own framework, Breit and Culbertson are incorrect. Clearly, Lerner's original two-person argument applies to each term in (12) independently of the others and consequently the equalization rule applies to the whole no matter how different the pairs of income deviations are in a test comparison.

Presumably Breit and Culbertson were led to (12) rather than the simpler, more symmetric, and relatively less constrained (4) by the wording of Lerner's original generalization:

Out of 100 million shifts away from an equalitarian distribution of income

in a large population, it could be expected that about 50 million would increase total satisfaction and about 50 million would diminish it. . . . The total increase in satisfaction received in the beneficial shifts would come to about 50 million times the shaded area  $G$  in the figure, while the total loss of satisfaction suffered in the harmful shifts would amount to about 50 million times the double-hatched area  $L$ . [p. 32]

Lerner's reference to the graphical analysis of his two-person theorem and his characterization of redistribution as 100 million income "shifts" certainly encourages interpretation of his large society as an  $n$ -person, two-utility function world of matched pairs with identical mean incomes. Breit and Culbertson at least consider the possibility of the matched pairs having different degrees of income inequality and the possibility of there being more than two utility functions. The general theorem we prove is not based on any special assumptions about matched pairs of incomes or utility functions.

## II. Lerner's Assumption of Ignorance

As shown above, Lerner's theorem is logically consistent when generalized to an  $n$  person society with any income distribution; hence future discussants of the theorem can concern themselves with more substantive issues. Since Lerner's theorem does not generate any potentially refutable predictions, its assumptions are the only source of substantive issues.

Of particular interest in the past has been Lerner's assumption of complete ignorance about which individual has which utility function, and his jump from that assumption to the hypothesis that each possible matching is equally likely. Friedman, p. 308, argues that Lerner's analysis is not rigorous because of his "equal ignorance" assumption.

While Samuelson, p. 175, does not agree with Friedman's argument, he feels enough doubt about the equal ignorance assumption to try to sidestep it through his proposed voting model reformulation of the problem. Breit and Culbertson, on the other hand,

seem comfortable with Lerner's essentially Bayesian approach. Similarly, up until this point, we have accepted the complete ignorance assumption and the hypothesis of equal probability in order to maintain the spirit of Lerner's original argument and to deal with the logical consistency raised by Breit and Culbertson. But, as we now show, the worrisome hypothesis of equal probability is not necessary for Lerner's conclusion.

Consider yet another route to (4). Social welfare,  $W$ , is assumed to consist of the sum of individual utilities. There are  $n!$  ways of allocating the  $n$  utility functions to the  $n$  incomes, thus yielding  $n!$  values for  $W$ , say  $W_1, \dots, W_{n!}$ . Under the Bayesian assumption each possibility is counted as equally probable and the mathematical expectation is

$$(13) \quad E(W) = \frac{1}{n!} \sum_{k=1}^{n!} W_k$$

This value is the same as that calculated from (4).

The need to make an assumption about the relative probabilities of various income-utility matchings arises because it has been assumed that society reacts to uncertainty by trying to maximize the expected value of social welfare. But now suppose that society adopts a "play safe" policy instead. That is, let it choose that income distribution which makes the worst of the possible values for  $W$  larger than the worst  $W$  which is associated with other income distributions. In other words, it adopts a *maximin* strategy.<sup>6</sup>

Under this strategy the government behaves as if the smallest  $W_k$  will result instead of  $E(W)$ . Since  $E(W)$  is the average of the  $W_k$ , it follows from (10) that

$$(14) \quad \begin{aligned} &W(\bar{Y}, \dots, \bar{Y}) \\ &\geq \min_k W_k(\bar{Y} + h_1, \dots, \bar{Y} + h_n) \end{aligned}$$

<sup>6</sup> A policy of play safe is familiar from the theory of games, and the present application might be interpreted as a "game against nature" in which nature chooses from the  $W_1, \dots, W_{n!}$  and the government selects the individual income shares.

Now when incomes are equally distributed, social welfare is unaffected by the distribution of utility functions, i.e., all  $W_k$  have the same value. Furthermore, that value is the one on the left-hand side of both (10) and (14). Thus (14) shows that an equal distribution of income is the maximin one and we find both expected value and *maximin* strategies leading to the same policy decision.<sup>7</sup>

This result definitely strengthens the appeal of Lerner's egalitarian advice. The play-safe strategy is the most pessimistic of policies which have any claim to rationality under uncertainty. In fact, other decision rules are usually considered because this strategy is unacceptably pessimistic (on the other hand, other strategies may turn out to be too optimistic in many situations, and the Bayesian approach is no exception). In this case, however, pessimistic decision makers and cautious but balanced optimistic decision makers would be in agreement.<sup>8</sup>

Even more reinforcement of the Lerner conclusion is possible. If the allocation of utility functions were known, the optimal distribution of income could be found with certainty and the maximum possible welfare attained. Moreover, this maximum value is the same no matter which way the functions

are allocated. Given any allocation of functions, the welfare from a given distribution of incomes can be calculated, and the shortfall from the maximum is the "regret" the government would have from having chosen this distribution instead of the best one. In particular, the even distribution yields a certain level of regret (the price paid for not knowing the actual allocation). It was shown previously that any other distribution *may* lead to a lower welfare level and hence it is associated with a greater maximum regret. Consequently, the even distribution policy also *minimizes* regret. Again this criterion does not, in general, lead to the same decision as does either the Bayesian or the maximin rule, but it happens to do so in this case. Thus the Lerner result satisfies yet another well-known criterion.<sup>9</sup>

Similar reinterpretations of Samuelson's voting result are possible. If each individual plays safe then he votes for an equal distribution of income, because he may get a lower income under any other distribution pattern. Alternatively, the best possible outcome for him under any one allocation is to have all the income. Consequently, the regrets associated with an equal division of income are all equal, while there is a larger regret for some allocation for any other distribution.<sup>10</sup> Notice also that under either of these new criteria, it makes no difference whether the voter is a risk taker or averter, and so the decision

<sup>7</sup> Figure 1 makes the result visually apparent in the two-person case. Starting from the position of equal incomes, it is clear that *any* redistribution *may* result in a loss of utility and therefore should be avoided under a play-safe strategy. The strategy simply gives no weight to the fact that there is also the possibility of a gain.

<sup>8</sup> This play safe policy is unrelated to probability in its basic philosophy, but it is instructive to give it a probability interpretation to contrast it with the Bayesian approach. The maximin strategy may be construed as that of maximizing the mathematical expectation of the  $W_k$  when the worst possibility from any income distribution (the smallest  $W_k$ ) gets assigned a probability of unity so that the other  $W_k$  get assigned zero probabilities. Given this interpretation, the play-safe strategy is clearly different from the "equal ignorance" postulate, and therefore it is not surprising that these two approaches under uncertainty frequently lead to different results. In this case, however, the symmetry with which the  $F_i$  enter  $W$  ensures that both approaches lead to symmetry in the optimal  $Y_j$ .

<sup>9</sup> Even so, these alternatives do not tackle a more fundamental objection. Lerner professes to be interested in unknown utility functions but, in fact, works with an unknown allocation of known functions. These are not the same though it makes no difference to the results when there is only a finite class of functions to choose from. Unfortunately, this is not the case. One of the advantages of Samuelson's model is that it keeps closer to Lerner's original intention in this respect.

<sup>10</sup> This argument assumes an "extensive" form of the voting procedure in which every possible permutation of the  $n$  individuals to the  $n$  income slots is distinguished. In practice, however, for administrative convenience, the voting forms might simply ask an individual to list his preferred  $n$  income shares in, say, descending order. Curiously enough, this mere change of form typically leads to a different *minimax* regret decision in which the voter will opt for a very unequal distribution, taking his chances on being allotted to a

under Samuelson's voting procedure applies to a wider class of cases than it does under Lerner's utility sum approach.<sup>11</sup>

large share. For example, a voter with a linear utility function would vote for shares 13/18, 4/18, 1/18, 0, . . . , 0. This radical change in the solution seems to be an application of the well-known proposition in minimax regret theory that the introduction of "irrelevant alternatives" may alter the decision, and the present case provides a good example of why many critics do not accept the *minimax* regret criterion as rational.

<sup>11</sup> On the other hand, the concavity assumption is used to prove the result in the case of maximizing the sum of utilities, and the decision might be wrong without the assumption. For example, if the curves in Figure 1 are relabelled *B* and *A* instead of *A* and *B*, then a sufficiently large redistribution is bound to result in a net gain. As always, concavity in what are essentially constrained maximization problems is sufficient but not necessary for the standard results. As Breit and Culbertson state, one marginal utility schedule can be increasing as long as the other decreases faster in the two-person case. Unfortunately, this weakening of the assumptions does not generalize in an interesting way,

since it is easy to see that all that can be allowed for *n* persons is that one marginal utility can be increasing and that not so fast as *any* of the others are diminishing.

#### REFERENCES

- W. Breit and W. P. Culbertson, Jr., "Distributional Equality and Aggregate Utility: Comment," *Amer. Econ. Rev.*, June 1970, 60, 435-41.
- M. Friedman, *Essays in Positive Economics*, Chicago 1953.
- A. P. Lerner, *The Economics of Control*, New York 1944.
- , "Distributional Equality and Aggregate Utility: Reply," *Amer. Econ. Rev.*, June 1970, 60, 442-43.
- I. M. D. Little, *A Critique of Welfare Economics*, London 1950.
- R. Musgrave, *The Theory of Public Finance*, New York 1959.
- P. A. Samuelson, "A. P. Lerner at Sixty," *Rev. Econ. Stud.*, June 1964, 31, 169-78.

# Distributional Equality and Aggregate Utility: Further Comment

By ROGER A. MCCAIN\*

Abba Lerner's formulation of the optimum division of income remains interesting although in some ways ambiguous. The comment by William Breit and William Culbertson in this *Review* is evidence of both points. There is still no formal proof of the theorem for an arbitrary number of persons, even if its outlines are clear enough. The "equal ignorance" assumption remains unclear. This note contains a proof of the "weak"<sup>1</sup> theorem for  $N$  persons. It uses a concept of ignorance derived from information theory.<sup>2</sup> Throughout this paper the term "income" is used in a somewhat special sense, following Lerner et al.; a sense somewhat akin to Gary Becker's "full income" (p. 497).<sup>3</sup> It is income evaluated on the assumption that the individual works the maximum feasible hours at maximum effort; purchases of (quantitative or qualitative) leisure are treated like any other purchases.<sup>4</sup> Moreover, it is assumed that we may discuss redistribution without considering the effect of resulting price changes on the utility-of-income schedules, (see Lerner (1944) pp. 23-24).

Suppose that the distribution of income by individual shares is given, that is, when  $Y_i$  is the income of individual  $i$ , and  $Y$  is total income  $Y_i/Y = y_i$  is a given constant for  $i = 1,$

$\dots, N$ . The  $y_i$  may be regarded as the probabilities that an incremental dollar of purchasing power goes to individual  $i$ .

Now, suppose it is known that each individual's utility function is drawn from a finite family  $f_j(Y_i)$ , for  $j = 1, \dots, J$ ;  $i = 1, \dots, N$ . The probability that individual  $i$ 's utility function is function  $j$  is  $p_{ij}$ . The expected utility of a dollar of purchasing power allocated to individual  $i$  is

$$(1) \quad E_i = \sum_{j=1}^J p_{ij} \frac{\partial f_j}{\partial Y_i}$$

The expected marginal social utility of the incremental dollar of purchasing power is

$$(2) \quad \begin{aligned} \sum_{i=1}^N E_i &= \sum_{i=1}^N y_i \sum_{j=1}^J p_{ij} \frac{\partial f_j}{\partial Y_i} \\ &= \sum_{i=1}^N \sum_{j=1}^J y_i p_{ij} \frac{\partial f_j}{\partial Y_i} \end{aligned}$$

The expected total utility is

$$(3) \quad U = \sum_{i=1}^N \sum_{j=1}^J p_{ij} f_j(y_i Y)$$

We wish to set the  $y_i$  so as to maximize  $U$ . Setting

$$(4) \quad Z = U + \lambda \left( \sum_{i=1}^N y_i - 1 \right),$$

and proceeding as usual, we obtain

$$(5) \quad \frac{\partial U}{\partial y_i} = -\lambda, \quad i = 1, \dots, N$$

as usual. That is

$$(6) \quad Y \sum_{j=1}^J p_{ij} \frac{\partial f_j}{\partial Y_i} = -\lambda$$

\* Assistant professor of economics, Western Washington State College.

<sup>1</sup> Like Lerner (1970) and unlike Breit and Culbertson, I find the strong theorem uninteresting because of its speculative character.

<sup>2</sup> This may be of interest because the legitimacy of the equal ignorance assumption has been denied. "... From complete ignorance nothing but complete ignorance can follow" (I. M. D. Little, p. 59). "From absolute ignorance we can derive nothing but absolute ignorance," (Graaff, p. 100fn). It is hoped that this paper may clarify the meaning of "complete ignorance."

<sup>3</sup> See also B. Mabry, pp. 213fn., 215fn.

<sup>4</sup> Compare Lerner (1944, p. 35).

The weighted sum of the marginal utilities must be equalized, as we would expect.<sup>5</sup>

Now, what does it mean to say we are "perfectly ignorant" about the distribution of the utility functions? Suppose that we establish telephone communication with Maxwell's Demon. We shall send him a message which states the identity of the person who is to receive the incremental dollar, and he will respond with the appropriate utility function. The probability that the "message sent" pertains to person  $i$  is  $q_{i.}$ , the probability that he will respond by designating function  $f_j$  is  $q_{.j}$ , and the probability that the message sent pertains to person  $i$  and the "message received" designates function  $j$  is  $q_{ij}$ . Here,  $q_{i.}$  and  $q_{.j}$  are the marginal probabilities,

$$q_{i.} = \sum_{j=1}^J q_{ij}$$

$$q_{.j} = \sum_{i=1}^N q_{ij}$$

Now,  $q_{i.}$  is the probability that the incremental dollar accrues to person  $i$ , that is,  $y_i$ . From the Demon's viewpoint,  $q_{.j}$  is the probability that he will designate function  $j$ , before he is told which person's utility function is needed. The expected mutual information of the two messages is (see Henri Theil, p. 34)

$$(7) \quad I = \sum_{i=1}^N \sum_{j=1}^J q_{ij} \log \frac{q_{ij}}{q_{i.} q_{.j}}$$

It is a measure of our knowledge of the interrelationship of persons  $i$  and utility functions  $j$  (compare Theil, pp. 31, 58). Perfect ignorance evidently implies that  $I=0$ , which is true only when the messages sent and received are stochastically independent, that is (see Theil, pp. 28, 31, 35),

<sup>5</sup> If  $\partial^2 f_j / \partial Y_i^2 < 0$  for all  $j$ , the absence of interpersonal interdependencies of utility assures that the second-order conditions are appropriate for a maximum. Indeed, all that is required is that the expected marginal utility of income diminishes; some of the  $\partial^2 f_j / \partial Y_i^2$  may be greater than zero consistently with this requirement. Breit and Culbertson's Figure 2 illustrates this possibility in the case  $N=J=2$ , p. 437.

$$(8) \quad q_{ij} = q_{i.} \cdot q_{.j}$$

Here, however,  $q_{i.} = y_i$  and moreover  $q_{ij} = y_i p_{ij}$ . Thus if  $I=0$ ,  $p_{ij} = q_{.j}$ . Equation (6) becomes

$$(9) \quad Y \sum_{j=1}^J q_{.j} \frac{\partial f_j}{\partial Y_i} = -\lambda$$

The weights must then be identical for all individuals. Thus to equalize the left-hand sides of (9) over  $i$  requires that the  $y_i$  be equal. Lerner's original argument for two persons was a special case of this argument, for a population of two persons. Unfortunately, the restriction to two persons introduces some confusing elements into the analysis.

Consider that, for  $N=2$  and  $J \leq N$ ,  $J$  must be either one or two. Now, there is nothing novel about the proposition that, when all utility functions are identical, equalitarian distribution maximizes total utility (assuming diminishing marginal utility). To illustrate his point, Lerner must assume, quite incidentally, that  $J=N$  (both are two) rather than  $J \leq N$ . Thus

$$(10) \quad \sum_{i=1}^N p_{ij} = 1,$$

as each utility function corresponds to just one person. Since  $p_{ij} = q_{.j}$ , this is

$$(11) \quad \sum_{i=1}^N q_{.j} = 1 = N q_{.j}; \text{ i.e., } p_{ij} = q_{.j} = \frac{1}{N}$$

The argument that equal distribution implies maximum expected utility is no less true if, for example,  $p_{11} = p_{21} = 1/4$ ;  $p_{12} = p_{22} = 3/4$ . What is required is that for given  $j$ ,  $p_{ij}$  be the same for all  $i$ . Then the expected utility functions,  $\sum_{j=1}^J p_{ij} f_j(Y_i)$  are identical for all individuals  $i$ .

By way of summary: if the expected utility functions are identical and characterized by diminishing marginal expected utility, then expected utility is at a maximum when incomes are equally distributed. The statement that we have zero information about the assignment of utility functions implies

that the expected utility functions are identical.

Unfortunately, the supposition that "each individual's utility function is drawn from a finite family" (above) is surely not so. The problem seems best resolved in the following way. Let  $S$  be the set of all conceivable utility functions. The Lerner-like theorem proved above is true of *each arbitrary finite subset* of  $S$ , if, and only if, all functions in  $S$  are differentiable and nonincreasing.<sup>6</sup> Now, the number of utility functions  $J$  can surely be no greater than the number of persons  $N$ ; hence the set of utility functions which people actually possess is a finite subset of  $S$ . Lerner's exact conclusion follows.

Lerner's theorem has no normative content, but, as Breit and Culbertson point out, some economists are evidently suspicious of it because they suspect that normative implications have been smuggled in somehow. There is no "illusionist's feat" (see p. 437), but Breit and Culbertson fail to explain why. It is hoped that the following comment will clarify the point.

What has been proved (accepting Lerner's assumptions 1.-5. (1944, pp. 23-28), and Breit and Culbertson, pp. 435-36) is the non-normative first term of the following syllogism:

1. "If it is desired to maximize the total satisfaction in a society, the rational procedure is to divide income on an equalitarian

<sup>6</sup> If even one utility function slopes upward, *that one* is a finite subset of  $S$  of which the Lerner-like theorem is false. The entire population may possess identical, upward-sloping utility functions. In that case the equalitarian conclusion is false. If a probability distribution over the entire set of utility functions can be defined, however, the argument may still be made; for we need only replace

$$\sum_{j=1}^J p_j \frac{\partial f_j}{\partial Y_i}$$

by the Lebesgue integral

$$\int_{-\infty}^{\infty} \phi \left( \frac{\partial f}{\partial Y_i} \right) \frac{\partial f}{\partial Y_i} d \frac{\partial f}{\partial Y_i},$$

where  $\phi$  is the density function of the marginal utility of money, assumed to be continuous almost everywhere (F. Riesz and B. Sz.-Nagy, chs. 1-2). This approach seems more in the spirit of Breit and Culbertson, pp. 437, 438, than Lerner.

basis" (Lerner (1944), p. 32; (1970), p. 422).

2. It is desired to maximize the total satisfaction in a society.

3. Income ought to be divided on an equalitarian basis.

The first premise is nonnormative; the second premise is equivalent to adopting the social welfare function<sup>7</sup>

$$(13) \quad W = U_1 + U_2 + \dots + U_N$$

Clearly, it is at that point that the ethical elements enter. The Lerner theorem has no more normative content than the "new welfare economics" developed along Bergson-Samuelson<sup>8</sup> lines; it differs from the Bergson-Samuelson welfare economics only in that it applies to a narrower class of social welfare functions.

If any value judgments have been smuggled in, they are the utilitarian values of the suspicious economists themselves. Lerner's equalitarian conclusion is like Kipling's Devil:<sup>9</sup>

"An' the Divil gave for answer, 'evict me if you can, sir,  
For I came in wid' the Donkey—on your Honour's invitation!'"

## REFERENCES

- G. Becker, "A Theory of the Allocation of Time," *Econ. J.*, Sept. 1965, 75, 493-517.  
A. Bergson, *Essays in Normative Economics*, Cambridge 1966.  
W. Breit and W. P. Culbertson, Jr., "Distributional Equality and Aggregate Utility: Comment," *Amer. Econ. Rev.*, June 1970, 60, 435-441.  
J. Graaff, *Theoretical Welfare Economics*, Cambridge 1967.  
J. Harsanyi, "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility," *J. Polit. Econ.* Aug. 1955, 63, 301-21.

<sup>7</sup> Note John Harsanyi, p. 314.

<sup>8</sup> Bergson, pp. 4-16; Samuelson, pp. 221-28.

<sup>9</sup> "The Legends of Evil: II," Verse, Definitive Edition (New York *Doubleday*, 1954 p. 355). Where Breit and Culbertson quote a classic, in translation, p. 440, I quote a neoromantic, in dialect. This must be symbolic of something.

- A. Lerner, *The Economics of Control*, New York 1944.
- , "Distributional Equality and Aggregate Utility: Reply," *Amer. Econ. Rev.* June 1970, 60, 442-43.
- I. M. D. Little, *Critique of Welfare Economics*, London 1958.
- B. Mabry, "An Analysis of Work and Other Constraints on Choices of Activities," *Western Econ. J.*, Sept. 1970, 8, 213-25.
- F. Riesz and B. Sz.-Nagy, *Functional Analysis*, New York 1955.
- P. Samuelson, *Foundations of Economic Analysis*, New York 1965.
- H. Theil, *Economics and Information Theory*, Amsterdam 1967.

# Distributional Equality and Aggregate Utility: Reply

By WILLIAM BREIT AND WILLIAM PATTON CULBERTSON, JR.\*

We take this opportunity afforded by the remarks of the present commentators to respond briefly not only to them but also to the earlier reply to our paper by Abba P. Lerner (1970). We are particularly delighted that so able an expositor of economic analysis considers that "The formulation of the argument for distributional equality by William Breit and William Culbertson is an improvement on that of *The Economics of Control* . . ." (1970, p. 442). It is also pleasant to be able to record our complete agreement with Lerner's revised statement of his famous proposition regarding the optimum division of income.

However it is necessary to point out that in his attempt to deny a switching of conclusions, Lerner has found it necessary to withdraw his bolder assertion that ". . . if it is desired to maximize the total satisfaction in a society, the rational procedure is to divide income on an equalitarian basis" (1944, p. 32, Lerner's italics). It is this proposition which Lerner was unable to prove given his assumptions. As we showed, the crucial assumption which Lerner's proof lacked is that each individual has an identical twin in the capacity to enjoy income. Unfortunately, Lerner seems to take this to mean that we have jettisoned his assumption regarding "equal ignorance." But this we would deny. Indeed, it was precisely to repair Friedman's restatement of Lerner that prompted us to submit our proof. For it is Milton Friedman's formulation, and not ours, which Lerner seems to be attacking. Our device of identical utility twins permitted us to allow for the possibility that redistribution may take place between persons with dissimilar utility functions, while at the same time resolving the problem that there may be some individuals with unique tastes. To arrive at this setting, we conducted our conceptual experiment.

Once this experiment was concluded, we again invoked the assumption of equal ignorance. That is, it is not necessary for us to know whether the utility function of any individual is greater than, equal to, or less than that of any other whom he may face in the redistributive process. However, since Lerner no longer wishes to claim his bolder proposition and since we have never taken issue with his much milder probability argument, we can only agree with his *restatement* of the bold assertion which now reads ". . . if it is desired to maximize the total satisfaction in a society, the rational procedure, in the absence of the knowledge that would enable us to equalize the marginal utilities, is to maximize the probable total satisfaction—i.e., to divide income on an equalitarian basis" (1970, p. 442, Lerner's italics). Whatever this sentence may have lost in Lernerian felicity, it has surely gained in precision.

While the comments by Maurice McManus, Gary Walton, and Richard Coffman (hereafter M-W-C) and Roger McCain deal with essentially similar issues, the papers are distinct enough to warrant separate comment. M-W-C ". . . offer a rigorous proof of Lerner's weak theorem for the *n*-person case" (p. 490). They then "contrast" their proof with our propositions for the purpose of disproving our alleged rejection of Lerner's theorem. But surely it should be evident that we have never rejected the weak theorem which Lerner presented in *The Economics of Control*. Therefore, to say that a "void" must be filled if "Lerner's theorem is to stand on a logical foundation" implies a rather serious confusion of the issues. We rejected Lerner's proof of his strong theorem and in subsequent comment he has rejected the strong theorem itself. This leaves Lerner holding only to his weak theorem, a proposition which is logically tenable as M-W-C neatly demonstrate.

However, it appears to us that a rather common error, an error to which we had pre-

\* Professor of economics and assistant professor of economics, University of Virginia, respectively.

viously called attention,<sup>1</sup> mars their attempt to extend Lerner's weak proposition. M-W-C claim to have found a result which "... definitely strengthens the appeal of Lerner's egalitarian advice" (p. 495). The result stems from supposing that society adopts a "play-safe" policy with regard to income redistribution. That is, rather than trying to maximize the expected value of individual utility, society adopts a *maximin* strategy which gives no weight to the possibility that redistribution results in a utility gain. Instead, the fact that "... *any* redistribution *may* result in a loss..." dictates that redistribution "... should be avoided under a play safe strategy" (fn. 7, their italics). M-W-C *prove* that this proposition is *consistent* with Lerner's egalitarian advice by referring to their Figure 1. And indeed, it is "visually apparent" that *starting from a position of equal incomes* the play-safe strategy indicates that equality be maintained since any redistribution *may* result in a loss. But this result is hardly identical to Lerner's utility sum approach. For suppose that incomes were not equally divided. Does the play-safe strategy suggest a redistribution to a more equal division as does the Lerner theorem? Hardly. Instead it suggests a status quo arrangement whatever the prevailing income distribution. Unfortunately, M-W-C have chosen to illustrate their proposition by referring to the single instance in which it is consistent with Lerner's egalitarian advice. Since the redistributive rule implicit in the play-safe strategy does not suggest movements toward income equality, it can hardly "strengthen" Lerner's theorem.

McCain's comments about our analysis at only two points. Our necessarily brief remarks will therefore be limited to these instances. 1) McCain finds "... the strong theorem uninteresting because of its specula-

tive character" (fn. 1). We find this comment singularly uninteresting because of its subjective character. 2) McCain's statement that Lerner's analysis involves no "illusionist's feat" is based on his belief that the Lerner theorem simply applies to a narrow class of Bergson-Samuelson social welfare functions. As he puts it, "If any value judgments have been smuggled in, they are the utilitarian values of the suspicious economists themselves" (p. 499). But this is precisely where we believe the Lerner redistribution argument has been left behind by recent developments in welfare economics. As we showed in our original paper it is possible to reconcile Lerner's approach with a voting model formulation in which, at the constitutional level of choice, unanimity would prevail with regard to egalitarian values. But that approach is very different in spirit from Lerner's since it is based on an individualistic rather than a utilitarian approach to redistribution. Lerner's demonstration of his egalitarian conclusion *appears* consistent with a Pareto optimal rule for redistribution only because probable satisfactions were maximized at his *equal income* starting point and a *status quo* arrangement was indicated. But once redistributions *toward* equality are recommended in an effort to maximize probable satisfactions, the illusion of ethical neutrality vanishes as the nonindividualistic ethic of Lerner's utility-sum approach becomes obvious.

#### REFERENCES

- W. Breit and W. P. Culbertson, Jr., "Distributional Equality and Aggregate Utility; Comment," *Amer. Econ. Rev.*, June 1970, 60, 435-41.
- A. P. Lerner, *The Economics of Control*, New York 1944.
- , "Distributional Equality and Aggregate Utility: Reply," *Amer. Econ. Rev.*, June 1970, 60, 442-43.

<sup>1</sup> See Breit and Culbertson (1970, p. 436, fn. 2) for a reference to de Jouvenel's confusion with regard to this point.

# ANNOUNCEMENT

## NOTICE TO ALL GRADUATE DEPARTMENTS

The December 1972 issue of the *Review* will carry the sixty-ninth list of doctoral dissertations in political economy in American universities and colleges. The list will specify doctoral degrees conferred during the academic year terminating June 1972. This announcement is an invitation to send us information for the preparation of the list. This announcement supercedes and replaces a letter which was sent annually from the managing editor's office.

The *Review* will publish in its December 1972 issue the names of those who will have been awarded the doctoral degree since June 1971, the titles of their dissertations, and, if possible, a brief (75-word) summary of the dissertation.

By June 30, please send us this information on 3×5 cards, conforming to the style shown below, one card for each individual. Please indicate by a classification number in the right-hand corner the field in which the thesis should be classified. The classification system is that used by the *Journal of Economic Literature* and printed in every issue.

Name: LAST NAME IN CAPS: First Name, Initial _____		JEL Classification No. _____
Institution Granting Degree: _____		
Degree Conferred (Ph.D. or D.B.A.) _____ Year _____		
Dissertation Title: _____		
Summary		
(75-word maximum, or first 75 words will be printed)		
Summary may be completed on back of this card or on new card which should be stapled to this.		

When degrees in economics are awarded under different names, such as Business Administration, Public Administration, or Industrial Relations, candidates in these fields whose training has been *primarily in economics* should be included.

# NOTES

## *Nominations for AEA Officers*

The Electoral College on March 24 chose Walter W. Heller as nominee for president-elect of the American Economic Association in the balloting to be held in the autumn of 1972. Other nominees (chosen by the nominating committee) are: for vice president (two to be elected), James S. Duesenberry, George Jaszi, Lawrence R. Klein, and Richard Ruggles; for member of the executive committee (two to be elected), Gary S. Becker, Robert L. Heilbroner, Robert J. Lampman, and Irma Adelman.

Under a change in the bylaws reported in the *Papers and Proceedings* of this Review, May 1971, additional candidates may be nominated by petition, delivered to the secretary by August 1, including signatures and addresses of not less than 6 percent of the membership of the Association for the office of president-elect and not less than 4 percent for each of the other offices. For the purpose of circulating petitions, address labels will be made available by the secretary at cost.

## *Notices to Members and Subscribers*

The American Economic Association is interested in purchasing certain issues of the *American Economic Review* and the *Journal of Economic Literature* published in 1970 and 1971. The issues desired and prices offered are:

*JEL*—March (\$3), June (\$3), and December (\$3), 1971.

*AER*—March (\$3), May (\$4), June (Regular issue, \$3; Supplement, \$1), and December (\$3), 1971; March (\$3) 1970.

The issues must be received in excellent condition before payment is made. Postage is to be paid by the shipper. Please send us a list of the above issues you desire to sell. We will advise you by return mail of the issues we approve of for purchase and will provide you with shipping instructions, packing slip, and address label. Authorization for purchase must be obtained from the AEA office before shipment is made. Correspondence should be addressed to: American Economic Association, 809 Oxford House, 1313 21st avenue South, Nashville, Tennessee 37212.

Members of the American Economic Association receive preregistration materials for the annual meetings together with the Association's publications. Normally subscribers do not receive preregistration materials because most subscriptions are held by institutions. In some cases, however, individual subscribers may wish to receive preregistration materials. These may be ob-

tained by writing to the secretary's office, Nashville, Tennessee.

The Asia Foundation has provided the American Economic Association with a fund to be used to assist students and visiting scholars from Asia who are studying in the United States or Canada to attend the annual meeting of the American Economic Association. The 1972 meeting will be held in Toronto, Canada, December 28-30. The maximum amount of money available to an individual applicant is \$150. Inquiries should be addressed to the American Economic Association, 1313 21st Avenue, South, Nashville, Tennessee 37212.

On October 26-27, 1972, the College of Business Administration of The University of South Carolina will sponsor a seminar on "Multiple Criteria Decision Making." The major goal of the seminar is to stimulate further research and practical developments in this area. The topics to be discussed include (1) the multiple objectives of the corporation in modern society, (2) the theory of macroeconomic stabilization policy, (3) multi-attributed utility theory, and (4) behavioral aspects of decision making under multiple goals. Among the speakers will be C. West Churchman, Yuji Ijiri, Kenneth MacCrimmon, Arthur Geoffron, and J. S. Dyer.

Persons interested in reading papers dealing with research and practical applications in the above and related areas are invited to submit a brief abstract, by June 1, 1972, to Professor Milan Zeleny, Department of Management Science, College of Business Administration, The University of South Carolina, Columbia, South Carolina 29208. Individuals interested in obtaining general information about the seminar should write to Professor Zeleny or Professor James L. Cochrane, Department of Economics, The University of South Carolina, Columbia, South Carolina 29208.

*Public Finance Quarterly* is a new journal for the study of the theory, policy, and institutions related to the allocation, distribution, and stabilization functions of the public sector. *PFQ* will begin publication January 1973. The Board of Editors invites original contributions in all areas of public finance. Style Sheets are available upon request.

Authors who wish to have their manuscripts considered for publication should send two copies to: Irving J. Goffman, Editor, *FPQ*, Department of Economics, University of Florida, Gainesville, Florida 32601.

Public finance specialists who wish to referee manuscripts should write the editor. Review copies of books should be sent to: J. Ronnie Davis, assistant editor, same address. Specialists who wish to review books should write the assistant editor.

The department of economics of the University of Wyoming, Laramie, 82070, will sponsor the first annual Bugas Summer School Program in Economics during the 1972 summer session. The program will include the offering of graduate courses and an Institute on the Development of Effective Health Manpower Planning Systems and Procedures. The visiting lecturers include: Professors John C. G. Boot, State University of New York at Buffalo; Charles E. McLure, Jr., Rice University; and James Quirk, California Institute of Technology. Dates for the Summer Session are June 12–August 4. The Institute is scheduled for August 14–18.

The *Journal of Economic Issues* announces the new location of its editorial offices. The *Journal* is a joint publication of the Association for Evolutionary Economics and the department of economics and Bureau of Economic and Business Research at Michigan State University, under the editorship of Professor Warren J. Samuels. The *Journal* is interested in receiving manuscripts comprising substantive contributions to the development and/or application of economic theory to problems and issues of economic policy and in the specialized fields of economics. A Style Sheet is available to authors. Please submit three copies to Professor W. J. Samuels, Department of Economics, Michigan State University, East Lansing, Michigan 48823.

It is proposed to establish an Association for the Economic Integration of North America. The association will be concerned with analysis of the benefits of particular measures of economic integration, especially among Canada, Mexico, and the United States; analysis of the costs of such integration; analysis of the equitable allocation of such costs; and dissemination of knowledge of current developments relating to such economic integration.

Persons interested in the objectives of such an association are invited to communicate with Professor Thomas M. Johnson, Department of General Business, Eastern Michigan University, Ypsilanti, Michigan 48104.

Aspecial summer program, Applied Urban Economics, will be held at Massachusetts Institute of Technology from August 21 to September 1, 1972. The program will feature classroom work and distinguished lecturers discussing the most recent theoretical and empirical economic findings in the area of urban affairs ranging from housing and transportation to the provision of government services, etc. For further information, write to Director of the Summer Session, Room E19-356, M.I.T., Cambridge, Massachusetts 02139.

The National Institute of Social and Behavioral Science will hold sessions for contributed papers at the 139th annual meeting of the American Association for the Advancement of Science, Dec. 26–31, 1972, in Washington, D.C. Economists interested in presenting

their research at these sessions may forward titles and abstracts of 300 words by Aug. 25 to Donald F. Ray, National Institute of Social and Behavioral Science, 863 Benjamin Franklin Station, Washington, D.C. 20044.

Suggested subjects might concern quasi-permanent extensions of a national incomes policy; questions of "overkill" and of priorities in environmental and welfare economics; problems of inflation relative to external and internal depreciation; monetary, credit, and fiscal policy; productivity and technological development; public sector productivity; the role of the factors of production in the distribution of average annual productivity increments; conditions for longer term equilibrium in international trade and finance; the *USSR* and East European trade with the West; neoinstitutionalism in contemporary economic thought; the goal of economic policy in reference to the unemployment rate; some measures for change in the adversary relationship between labor and management; and topics in interdisciplinary studies.

### Deaths

Nicholas C. Anagnos, Bethesda, Maryland.

Tranquilino B. (Frank) Aquino, retired Federal Trade Commission economist, Falls Church, Virginia, Nov. 6, 1971.

Charles E. Barrett, professor of economics, Industrial College of the Armed Forces, Washington, Dec. 1971.

Gene L. Chapin, associate professor, department of economics, Ohio University Mar. 7, 1972.

John Cornell, Bradenton, Florida, Jan. 15, 1972.

Charles E. Ferguson, professor, Texas A&M University, Jan. 14, 1972.

A. L. Fleming, department of economics, Salisbury State College, Feb. 29, 1972

Herbert A. Howard, economics department, School of Business, Auburn University, Mar. 10, 1972.

Almon T. Mace, chairman, department of business administration and economics, Madison College, Mar. 7, 1972.

Henry C. Murphy, Washington, D.C., Jan. 1972.

S. H. Nerlove, professor emeritus, University of Chicago and retired professor in residence at University of California, Los Angeles, Feb. 13, 1972.

Lynn A. Stiles, vice-president and economist, Federal Reserve Bank of Chicago, Mar. 19, 1972.

### Retirements

Chandler Morse, professor emeritus, Cornell University, July 1, 1971.

A. J. Penz, professor emeritus, University of Alabama, June 1, 1971.

### Visiting Foreign Scholars

Ragnar Bentzel, Uppsala University, Sweden: visiting professor of economics, University of Florida, Mar. 1972, spring quarter.

Jacques H. Dreze, Universite Catholique de Louvain, Belgium: visiting scholar, Cornell University, 1971-76.  
 Jacquelin L. Hodgson, University of the Americas, Mexico: visiting professor of economics, University of Florida, 1971-72.

Alan A. Tait, University of Strathclyde, Glasgow: visiting scholar, fiscal affairs department, International Monetary Fund, Jan.-Mar. 1972.

### *Promotions*

Douglas K. Aide: associate professor, department of economics, Ohio University, Sept. 1, 1972.

William J. Barger: assistant professor of economics, University of Southern California, fall 1971.

Kelly J. Black: associate professor, Chico State College.

Harvey Botwin: associate professor of economics, Pitzer College, Associated Colleges at Claremont, Sept. 1972.

Boyd M. Buxton: associate professor of agricultural and applied economics, University of Minnesota and head, North Central Field Group, Economic Research Service, USDA.

John J. Casson: associate economist and assistant manager, Brown Brothers, Harriman & Co., New York.

Elmer G. Dickson: associate professor, Chico State College.

Kenneth E. Egertson: associate professor and extension economist, marketing, agricultural and applied economics and extension service, University of Minnesota.

Peter Gordon: assistant professor of economics and urban and regional planning, department of economics, fall 1971.

Bartell Jensen: professor, department of economics, Utah State University.

George D. Johnson: associate professor, Chico State College.

Arthur Kraft, associate professor of quantitative methods, Ohio University, June 1, 1972.

Thomas G. Macbeth: professor of economics, Lowell Technological Institute, Feb. 1972.

John O. Mason: associate professor of accounting, University of Alabama, Aug. 22, 1971.

Bruce Newling: associate professor of economics, City College of the City University of New York.

Albert V. Niemi, Jr.: associate professor of economics, University of Georgia.

William O'Keefe: director of administration, Center for Naval Analyses, University of Rochester, Feb. 1972.

L. Andrew Potemra: associate professor, department of economics, Ohio University, Sept. 1, 1972.

Morris Silver: professor of economics, City College of the City University of New York.

Roger M. Swagler: assistant professor of economics, Drake University, Sept. 1971.

Simón Teitel: adjunct associate professor, department of economics, Catholic University of America, Sept. 1, 1971.

### *Administrative Appointments*

Raymond R. Beneke: acting head, department of economics, Iowa State University, Feb. 14, 1972.

Clement H. Donovan: director of African studies, University of Florida, Sept. 1971.

B. Delwroth Gardner: head, department of economics, Utah State University.

Oscar R. Goodman: chairman, finance department, Roosevelt University, Sept. 1, 1972.

Charles W. Howe: chairman, department of economics, University of Colorado, Feb. 1972.

Richard Hyse: chairman, department of economics, New York State University College at Oswego.

David Kassing, Clarkson Institute of Technology: executive vice president, Center for Naval Analyses, University of Rochester, Jan. 1972.

Barbara H. Kehrer, Fisk University: research associate; director, department of social and economic research, American Medical Association, Dec. 1971.

Adamantios Pepelasis: head, department of economics, Virginia Polytechnic Institute and State University, 1971-72.

Paul E. Roberts: director of MBA program, College of Business Administration, University of Florida, Sept. 1971.

### *New Appointments*

Kenneth R. Biederman: assistant professor, department of economics, Georgetown University, Sept. 1971.

Bradley B. Billings: assistant professor, department of economics, Georgetown University, Sept. 1971.

Roy Blough, Columbia University: visiting professor of finance, University of Florida, winter quarter 1971-72.

Tuvia Blumenthal, Hebrew University: senior lecturer in economics and developing countries, Tel Aviv University.

Clark R. Burbee, University of Minnesota: agricultural economist, Poultry Group, MED, Economic Research Service, USDA.

David C. Campbell, University of California, Berkeley: assistant professor, department of economics, University of Idaho, Jan. 1972.

Anne Carter visiting lecturer in economics, Brandeis University, 1971-72.

A. John Deboer, University of Minnesota: lecturer, department of economics, Faculty of Agriculture, University of Queensland, Brisbane, Australia.

Robert Evans: professor of economics, Brandeis University, Sept. 1971.

Phillip Friedman, Massachusetts Institute of Technology: assistant professor of economics, University of Florida, spring 1972.

Paul Gatons, Georgia State University: assistant professor of economics, department of economics and finance, Louisiana Tech University, Sept. 1971.

Allan Glubok, Colorado State University: associate professor of finance, department of economics and finance, Louisiana Tech University, Sept. 1971.

Steven D. Gold: instructor in economics, Drake University, Sept. 1971.

Karl D. Gregory: associate professor, School of Economics and Management, Oakland University, Aug. 15, 1971.

Klaus D. Grimm: research associate, U.S. Department of Labor, Region LX, San Francisco.

Morley Gunderson, University of Wisconsin: assistant professor of economics, Center for Industrial Relations, University of Toronto, Sept. 1971.

Charles E. McConnel: assistant professor of economics, Alfred University, Sept. 1971.

Allan B. Mandelstamm, Michigan State University: visiting professor of economics, University of Florida, winter quarter, Dec. 1971.

M. Herschel Mann: assistant professor, University of Alabama, Aug. 22, 1971.

Neil B. Murphy, Federal Deposit Insurance Corporation: professor of finance, College of Business Administration, University of Maine.

George W. Parker, University of Iowa: assistant professor of economics, Mississippi State University, Sept. 1971.

Robert W. Pearson: assistant professor of economics, Goucher College, 1971-72.

Kenneth E. Raske, Northern Illinois University: research associate, department of social and economic research, American Medical Association, June 1971.

John W. Schamper, University of Wisconsin: research associate, agricultural and applied economics, University of Minnesota.

Rainer Schickele: visiting professor, agricultural and applied economics, University of Minnesota, Jan.-June 1972.

Sandra G. Schickele: assistant professor of economics, Pitzer College, Associated Colleges at Claremont, Sept. 1971.

William C. Stubblebine, Claremont Men's College: visiting associate professor of economics, Virginia Polytechnic Institute and State University, winter 1972.

John Tuccillo: assistant professor, department of economics, Georgetown University, Sept. 1971.

Dudley Wallace, North Carolina State University: visiting professor of econometrics, Virginia Polytechnic Institute and State University, 1972.

James S. Weshow: visiting professor of economics, University of Florida, Dec. 1971 winter quarter.

Henry W. Zaretsky, University of California, Davis: research associate, department of social and economic research, American Medical Association, Sept. 1971.

### *Leaves for Special Appointments*

Howard S. Ellis, University of California Berkeley: visiting professor, University of Wisconsin, Milwaukee, spring semester 1972.

Charles W. Fristoe, University of Florida: visiting professor, University of the Americas, Sept. 1971, fall and winter quarters.

Richard D. Haas, University of Georgia: visiting lecturer, University of Nottingham, England, 1971-72.

Jerome W. Hammond, University of Minnesota: party chief, Minnesota-U.S. AID Project, Tunis, Tunisia.

Christopher I. Higgins, Australian Treasury: consultant in economic modelling, U.N. Computing Research Centre, Bratislava, Czechoslovakia, May 1972.

Harald R. Jensen, University of Minnesota: Office of Agriculture and Fisheries, Bureau for Technical Assistance, Department of State, Washington.

Louis J. Junker, Western Michigan University: Economic Planning Unit of Ministry of Economic Planning and Development, UNESCO, Mauritius.

Jeffrey B. Nugent, University of Southern California: United Nations, Beirut, Lebanon, Sept. 1, 1971-June 30, 1972.

Ronald J. Vogel, University of Florida: Brookings Fellowship, Department of Health, Education, and Welfare, July 1971-June 1972.

Pascal J. Wick, University of Minnesota: Agency for International Development, Tunisia, 1971-73.

### *Resignations*

Anthony Cephalas, Oakland University: Center for Economic Research, Athens, Greece, Aug. 14, 1971.

M. June Flanders, Purdue University: Tel Aviv University.

Jerome M. Stam, University of Minnesota: Economic Research Service, U.S. Department of Agriculture.

Paul Van Moeseke, Iowa State University: University of Louvain, Belgium, Nov. 1, 1971.

### *Miscellaneous*

Elise G. Jancura, Cleveland State University: editorial board, *The Woman CPA*.

E. Bryant Phillips, University of Southern California: president of the Omicron Delta Epsilon, Jan. 1972.

#### NOTE TO DEPARTMENTAL SECRETARIES AND EXECUTIVE OFFICERS

When sending information to the *Review* for inclusion in the Notes Section, please use the following style:

A. Please use the following categories:

- |   |   |
|---|---|
| 1—Deaths  | 6—New Appointments                                  |
| 2—Retirements                                   | 7—Leaves for Special Appointments (NOT Sabbaticals) |
| 3—Foreign Scholars (visiting the USA or Canada) | 8—Resignations                                      |
| 4—Promotions                                    | 9—Miscellaneous                                     |
| 5—Administrative Appointments                   |   |

B. Please give the name of the individual (SMITH, John W.), his present place of employment or enrollment: his new title (if any), his next place of employment (if known or if changed), and the date at which the change will occur.

C. Type each item on a separate 3×5 card, and please do not send public relations releases.

D. The closing dates for each issue are as follows: *March*, November 1, *June*, February 1, *September*, May 1, *December*, August 1.

This announcement supersedes and replaces a letter which was sent annually from the managing editor's office.

---

# The American Economic Review

Published by the  
American Economic Association

The American Economic Association was organized in 1886 to promote the study of political economy in the United States and to advance the science of economics.

Editorial Board: *James M. Buchanan*, *Robert H. Frank*, *John H. Garvey*, *James A. Hines*, *Robert A. J. O'Connell*, *James M. Poterba*, *Robert C. Shiller*, *James H. Stock*, *John H. Williams*, *James J. Heckman*, *James D. Watson*, *James H. Poterba*, *Robert C. Shiller*, *James H. Stock*, *John H. Williams*, *James J. Heckman*, *James D. Watson*.

Editorial Board: *James M. Buchanan*, *Robert H. Frank*, *John H. Garvey*, *James A. Hines*, *Robert A. J. O'Connell*, *James M. Poterba*, *Robert C. Shiller*, *James H. Stock*, *John H. Williams*, *James J. Heckman*, *James D. Watson*, *James H. Poterba*, *Robert C. Shiller*, *James H. Stock*, *John H. Williams*, *James J. Heckman*, *James D. Watson*.

Editorial Board: *James M. Buchanan*, *Robert H. Frank*, *John H. Garvey*, *James A. Hines*, *Robert A. J. O'Connell*, *James M. Poterba*, *Robert C. Shiller*, *James H. Stock*, *John H. Williams*, *James J. Heckman*, *James D. Watson*, *James H. Poterba*, *Robert C. Shiller*, *James H. Stock*, *John H. Williams*, *James J. Heckman*, *James D. Watson*.

Editorial Board: *James M. Buchanan*, *Robert H. Frank*, *John H. Garvey*, *James A. Hines*, *Robert A. J. O'Connell*, *James M. Poterba*, *Robert C. Shiller*, *James H. Stock*, *John H. Williams*, *James J. Heckman*, *James D. Watson*, *James H. Poterba*, *Robert C. Shiller*, *James H. Stock*, *John H. Williams*, *James J. Heckman*, *James D. Watson*.

Editorial Board: *James M. Buchanan*, *Robert H. Frank*, *John H. Garvey*, *James A. Hines*, *Robert A. J. O'Connell*, *James M. Poterba*, *Robert C. Shiller*, *James H. Stock*, *John H. Williams*, *James J. Heckman*, *James D. Watson*.

Editorial Board: *James M. Buchanan*, *Robert H. Frank*, *John H. Garvey*, *James A. Hines*, *Robert A. J. O'Connell*, *James M. Poterba*, *Robert C. Shiller*, *James H. Stock*, *John H. Williams*, *James J. Heckman*, *James D. Watson*, *James H. Poterba*, *Robert C. Shiller*, *James H. Stock*, *John H. Williams*, *James J. Heckman*, *James D. Watson*.

Editorial Board: *James M. Buchanan*, *Robert H. Frank*, *John H. Garvey*, *James A. Hines*, *Robert A. J. O'Connell*, *James M. Poterba*, *Robert C. Shiller*, *James H. Stock*, *John H. Williams*, *James J. Heckman*, *James D. Watson*, *James H. Poterba*, *Robert C. Shiller*, *James H. Stock*, *John H. Williams*, *James J. Heckman*, *James D. Watson*.

Editorial Board: *James M. Buchanan*, *Robert H. Frank*, *John H. Garvey*, *James A. Hines*, *Robert A. J. O'Connell*, *James M. Poterba*, *Robert C. Shiller*, *James H. Stock*, *John H. Williams*, *James J. Heckman*, *James D. Watson*, *James H. Poterba*, *Robert C. Shiller*, *James H. Stock*, *John H. Williams*, *James J. Heckman*, *James D. Watson*.

Editorial Board: *James M. Buchanan*, *Robert H. Frank*, *John H. Garvey*, *James A. Hines*, *Robert A. J. O'Connell*, *James M. Poterba*, *Robert C. Shiller*, *James H. Stock*, *John H. Williams*, *James J. Heckman*, *James D. Watson*, *James H. Poterba*, *Robert C. Shiller*, *James H. Stock*, *John H. Williams*, *James J. Heckman*, *James D. Watson*.

The American Economic Review is published by the American Economic Association, which was organized in 1886 to promote the study of political economy in the United States and to advance the science of economics. The Association is composed of economists from all parts of the United States and from other countries. The Review is a journal of the Association and is published quarterly. It contains articles on a wide range of economic subjects, including theory, policy, and statistics. The Review is one of the leading journals in the field of economics and is read by economists throughout the world.

Volume 100, Number 1, January 2010

# THE AMERICAN ECONOMIC ASSOCIATION

Founded in 1885

• Published at George Banta Co., Inc., Menasha, Wisconsin.

• THE AMERICAN ECONOMIC REVIEW, including four quarterly numbers, the *Proceedings* of the annual meetings, and *Directory* and Supplements, is published by the American Economic Association and is sent to all members five times a year, in March, May, June, September, and December.

• Membership dues of the Association are \$20.00 a year, which includes a year's subscription to both the *American Economic Review* and the *Journal of Economic Literature*. Subscriptions by nonmembers are \$30.00 a year, and only subscriptions to both publications will be accepted. Single copies of the *Review* and *Journal* are \$4.00 each. Each order for copies of either publication must also include a \$.50 per order service charge. Orders should be sent to the Secretary's office, Nashville, Tennessee.

• Correspondence relating to the *Papers and Proceedings*, the *Directory*, advertising, permission to quote, business matters, subscriptions, membership and changes of address may be sent to the secretary, Rendigs Fels, 1313 21st Avenue, South, Nashville, Tennessee 37212. To be effective, notice of change of address must reach the secretary by the 1st of the month previous to the month of publication. The Association's publications are mailed by second class and are not forwardable by the Post Office.

• Second-class postage paid at Nashville, Tennessee and at additional mailing offices. Printed in U.S.A.

## Officers

### *President*

JOHN KENNETH GALBRAITH  
Harvard University

### *President-Elect*

KENNETH ARROW  
Harvard University

### *Vice-Presidents*

HENDRIK S. HOUTHAKKER  
Harvard University  
ARTHUR M. OKUN  
Brookings Institution

### *Secretary-Treasurer and Editor of Proceedings*

RENDIGS FELS  
Vanderbilt University

### *Managing Editor of The American Economic Review*

GEORGE H. BORTS  
Brown University

### *Managing Editor of The Journal of Economic Literature*

MARK PERLMAN  
University of Pittsburgh

## Executive Committee

### *Elected Members of the Executive Committee*

ROBERT DORFMAN  
Harvard University  
ARNOLD C. HARBERGER  
University of Chicago  
ROBERT EISNER  
Northwestern University  
JOHN R. MEYER  
Yale University  
GUY HENDERSON ORCUTT  
Yale University  
JOSEPH A. PECHMAN  
Brookings Institution

### *Ex Officio Members*

WASSILY LEONTIEF  
Harvard University  
JAMES TOBIN  
Yale University

## TJALLING C. KOOPMANS

DISTINGUISHED FELLOW

1971

In designating Tjalling Koopmans a Distinguished Fellow, we honor a revolutionary leader whose movement has peacefully but profoundly transformed economic theory and method in one generation. He has pursued with single-minded dedication the goal of logical precision in quantitative economics. His pursuit has attracted, encouraged, and inspired a growing company of quantitatively minded economists, many of them his own colleagues and students. He played central roles in elucidating the principles of identification and estimation of econometric models, and in formulating the activity analysis model of production, with its fruitful implications for linear programming and for study of decentralization and efficiency. More recently he has been making characteristically fundamental contributions to the analysis of optimal growth and efficient allocation of resources over time. In seminar and in print, by achievement and by example, he inspires us, young and old, with faith in the scientific destiny of our discipline.



*Tjalling. Koopman*

# The 1972 Report of the President's Council of Economic Advisers: Inflation and Unemployment

By EDGAR L. FEIGE\*

The optimistically green clad 1971 Economic Report of the President (best known for its rosy \$1,065 billion *GNP* forecast) has been superseded by a more sober report for 1972 which is appropriately draped in purple. The economic story for the past year reveals that realized *GNP* fell \$18 billion short of the Council's forecast target; unemployment stubbornly held at a 6 percent rate; the \$11.8 billion fiscal year forecast deficit turned into a \$38.8 billion deficit; the balance of payments continued to deteriorate; and the inflation rate remained higher than anticipated. The poor showing on the economic scoreboard for 1971 coupled with the enormous stakes of winning the 1972 political ballgame led a disappointed and embarrassed Administration to scrap their game plan and punt.

When economic realities fall grossly short of political wishful thinking (even when couched in the guise of econometric forecasts), either the goals must be candidly pared down or dramatic action must be taken to improve the state of the economy. Facing an election year, the Administration chose to adopt both techniques to narrow the gap.

With respect to goals, the 1971 report had envisioned "a path that would see the unemployment rate reduced to the 4½ percent zone by the second quarter of 1972 and the inflation rate, . . . declining to ap-

proach the 3 percent range at that time" (p. 78). The 1972 report, in sharp contrast, predicts a fall in the rate of unemployment to the "neighborhood" of 5 percent *by year end* (no average rate for the year is reported), and a 3¼ percent increase in prices *during* the year.

By way of action, the Administration responded to the disappointing first half of 1971 by introducing the August 15 New Economic Policy (*NEP*). Both in conception and execution, the *NEP* package was a bold stroke which departed significantly from earlier Administration policy pronouncements. The Council's Report demonstrates that the customary lag between the generation of economic knowledge and the application of that knowledge to public policy has been greatly reduced. The *NEP* package goes considerably beyond the single issue recommendations of many private economists and it mirrors in concrete terms many of the ambiguities and controversies which presently exist within the profession. While professional economists can boast that the "consensus forecast" for *GNP* was right on target, there is unlikely to be as clear a consensus in the profession on the most effective means for simultaneously reducing the balance-of-payments deficit, the rate of unemployment, and the rate of inflation. The Council's highly qualified justifications for the Administration's actions reflect an awareness of recent contributions to economics and, perhaps more importantly, an appreciation of the tentative

\* Associate professor of economics, University of Wisconsin. I am indebted for helpful comments on earlier drafts to Glen Cain, Arthur Goldberger, Lee Hansen, Irene Lurie, Donald Nichols, and Leanna Stiefel.

nature of those contributions. The Council argues:

The problems of managing fiscal policy or monetary policy or both have apparently been underestimated. It may well be that more has been promised than can be delivered with existing knowledge and instruments. Certainly there is need for much additional research. But if the question is not one of keeping the economy on a narrowly-defined path but one of avoiding violent aberrations like the one that began in 1965, our tools are probably adequate, and the problem is more the national will than the techniques of economics and economic policy. [p. 112]

The original Nixon game plan was to reduce inflationary expectations by moderating and stabilizing both monetary and fiscal instruments in the hope of gradually reducing inflation at the cost of a minimum increase in unemployment. In stark contrast to the Administration's intentions, monetary growth was highly unstable, fiscal policy produced large deficits, inflation was relatively unyielding, and unemployment steadily climbed. The monetarist position which seemed to receive considerable attention by the Administration did not achieve its hoped-for results and by mid-1971 the national will clamored for a more aggressive policy. While monetarists will argue with considerable justification that their position did not receive an adequate test, the fact remains that the continued rise in prices and the emergence of high rates of unemployment made a continuation of the original game plan politically unviable (see A. James Meigs).

What, then, were the problems confronting the professionals and the Administration, and how were these problems dealt with?

The Council candidly acknowledges the policy dilemmas that it faced in 1971:

... the American economy was beset by a conflict among four objectives—faster growth, higher employment, greater

price stability, and a more balanced external position. The danger was that steps to speed up growth and boost employment by expanding demand would worsen both the inflation and the balance-of-payments deficit. . . . The measures taken last year did not *eliminate* potential conflicts among these goals. . . . But the area of consistency among the objectives has been greatly widened. [p. 101]

How was the area of greater consistency achieved? On the international front, the Administration suspended the convertibility of dollars into gold and other international reserve assets and imposed a temporary 10 percent import surcharge. While these measures did not bring about the system of flexible exchange rates academic economists had hoped for, they did produce a long overdue revaluation of currencies and thus temporarily reduced the pressure on the balance of payments. The imposition of the import surcharge can be criticized for creating a climate conducive to retaliation and perhaps strengthening protectionist sentiment. Moreover, it may have inhibited a more drastic realignment of currency valuations. *Ex post*, however, it appears that the surcharge was a politically expedient means of speeding the necessary currency revaluations, and more importantly, of bringing the international monetary community to the acceptance of a broader band of permissible fluctuations in exchange rates. While these measures are no substitute for longer term reforms of international monetary arrangements, they do move in the direction of reducing the balance-of-payments problem and thus allow greater freedom to deal with the domestic problems of inflation and unemployment.

To deal with inflation, the Administration adopted an "incomes policy" which ironically had found its severest critics, prior to August 15, within the Administration itself. The Council argued that the

suddenness of its action was one of the more important ingredients in its forecasted success since it presumably precluded "anticipatory price and wage increases."

Prior to August 15, the Administration had adopted the position of many economists that direct controls are not likely to be effective, and indeed, to the extent that they are, can be positively detrimental because of their allocative and distributive consequences. As inflationary pressure continued in the face of growing unemployment, the case for an incomes policy became more widespread among professional economists. With the benefit of hindsight, it is now argued that while the wage-price freeze may have been better late than never, it was unambiguously late. The justification for an incomes policy was found in the Phillips curve literature which recently has placed considerable emphasis on the role of expectations. This literature argued that a slowing of the inflation rate during periods of high unemployment would reduce the expectations of inflation and this, in turn, would shift the short-run Phillips curve to the left, improving the short-run tradeoff between inflation and unemployment. In the long run, the tradeoff would vanish and the unemployment rate would return to its "natural" level (see Milton Friedman). It was thus only a short jump to the proposition that an exogenous shock to "inflationary psychology" could improve the short-run tradeoff between unemployment and inflation.

The Economic Report places the major burden of justification for the wage-price freeze in particular, and for the *NEP* package more generally, on their effects on reducing inflationary expectations. It argues that the control system "... can generate the *expectation* of reasonable price stability that is essential to the *achievement* of reasonable price stability"

(p. 27), and that the control system is "... intended to stimulate spending and employment by reducing the inflation-anxiety of consumers and businessmen" (p. 24). Whether such claims will be validated by experience is by no means obvious. However, a critique of the Council's predictions must take account of the plethora of qualifications which are sprinkled around every forecast, and is made difficult by the Council's studious avoidance of an explicit model.

The Council's position raises two inter-related questions. First, will the wage-price freeze and Phase II in fact succeed in reducing inflationary expectations? Second, will such a reduction in expectations bring about the specific outcomes predicted by the Council, namely reduced inflation, increased spending, and lower unemployment? There are numerous considerations which suggest a negative response to both questions. Price expectations depend not only on actual inflationary experience but also on the extent to which economic units believe their expectations will be validated by government actions. To the extent that monetary and fiscal policies are viewed as highly expansionary, the imposition of a temporary freeze is unlikely to persuade economic units that the reduction in the actual rate of inflation brought about by a temporary freeze is going to be long lasting. The Council recognizes this effect in its discussion of automobile purchases by arguing that "The increase in sales was also influenced by the expectation of a price rise after the freeze" (p. 25), implying a post-freeze result with higher prices and lower sales.

But if automobile prices are expected to rise, what would dampen similar expectations of price rises of other goods and services? The Council asserts that "The control system... is meant to assist market forces that would be working to

hold down inflation; it is not meant to resist market forces working to accelerate inflation" (p. 27). One suspects that the operational implications of this tortuous distinction may escape not only professional economists but more importantly the price board bureaucrats, who, given the Council's figures, were expected to answer an average of 24,000 daily inquiries during the months following the initiation of Phase II.

An apocryphal justification for Phase II is that if you can fool some of the people all of the time and all of the people some of the time, that puts the odds heavily in your favor. Unfortunately, an incomes policy which sets guidelines for price increases can actually have the effect of alerting economic units to *raise* prices to the guideline limits.

The Administration, having previously articulated a strong case against direct controls, hoped that by selecting an incomes policy which was both temporary and limited in scope, it would achieve a reduction in inflationary expectations with minimal misallocation consequences. It seems equally likely that the wage-price policy will have a perverse effect on the price level and tend to stifle the allocative transition from a guns to butter economy.

The Administration's response to the unemployment situation was an apparently aggressive fiscal policy coupled with an attempt to understate the magnitude of the unemployment problem. Once again, the Administration relied heavily on recent developments in the professional economics literature. For example, the justification for a higher acceptable unemployment rate was found in the changing composition of the labor force (see George Perry). Recent Administration pronouncements confirm what is already implicit in the Council's Report, namely, that the full employment target has been shifted from the previously accepted 4 percent

unemployment rate to a 5 percent unemployment rate.

The Council, along with many professional economists, has become sensitive to the macro-economists' warning that aggregates are not homogenous. The retreat to micro-economic explanations of difficult macro-economic phenomena seems to be highly correlated with the experience of continuous inflation in the face of higher than average rates of unemployment. Thus in the late 1950's and early 1960's, changes in the composition of the *demand* for labor became fashionable explanations for simultaneous unemployment and inflation. Rapid automation and structural shifts became nominees for the villain's role in producing poor macro-predictions. In the 1970's, the burden has shifted to the shoulders of a changing composition of the *supply* of labor. It is now argued that the labor force contains a higher proportion of teenagers and women who have historically experienced above average rates of unemployment. Thus, an aggregate unemployment rate of 4 percent with the old age-sex labor force composition is now comparable to a higher aggregate rate given the new composition. However, the converse argument also holds when labor force composition is analyzed from the perspective of educational attainment level. The present labor force also contains a larger percentage of more highly educated persons who historically have had lower unemployment rates. Thus, the education composition effect would argue that an aggregate unemployment rate of 4 percent could now be more easily attained. The Council's Report focuses attention on the age-sex composition effects and seems to conclude that general expansion policies will have a lower bang per buck in reducing the unemployment rate. It therefore relies heavily on programs and policies that will reduce "transitional" unemployment. Among these are improved job information

services, retraining programs, and a reduction of the minimum wage for teenagers.

With respect to expansive policies the Council adopted the view that:

The establishment of the direct wage-price controls created room for some more expansive measures, because it provided a certain degree of protection against both the fact and the expectation of inflation. This situation had to be approached with caution, because excessive expansion could make the price-wage control system unworkable. Still there could be no doubt that the tolerable rate of expansion had been increased.

[p. 69]

If 1971 revealed that President Nixon had made his debut as a Keynesian, 1972 heralded him as an ardent Keynesian with a vengeance. "Some more expansive measures" became a startling \$38.8 billion deficit for fiscal 1972, implying a full employment (4 percent) deficit of \$8.1 billion. If, however, one accepts the Administration's view that full employment is now characterized by a 5 percent unemployment rate, the appropriately redefined "full employment" deficit is closer to \$23.5 billion. Moreover, when the budgeted outlays for fiscal 1972 are compared to actual expenditures undertaken in the first half of the fiscal year, it becomes apparent that an enormous upsurge in expenditures must be forthcoming in the second half of fiscal 1972. The timing and magnitude of these federal expenditures appear to be aimed more at the November elections than toward an efficient allocation of expenditures.

Given what must be regarded as a large full employment deficit, it is puzzling that the unemployment rate remained so stubborn. One explanation may be found in a closer analysis of the deficit's components. The Administration, recognizing that temporary tax cuts on expenditures have larger multiplier effects than temporary tax cuts on income, provided businesses

with a tax credit for new investment and repealed the 7 percent excise tax on automobiles. While the excise tax cut clearly led to a spurt in automobile purchases, the expected multiplier effects of the investment tax credit were blunted both by excess capacity and by congressional actions which eliminated the two-tier form of credit proposed by the Council. Moreover, a large part of the federal outlays consisted of increases in unemployment compensation, social security payments, and military pay increases which are not likely to have the full employment consequences of direct federal expenditures on goods and services (see Leonard Silk). Thus while the stimulating effects of fiscal policy may be less than those anticipated by simply gauging the magnitude of the deficit, it is likely that the size of the deficit will work against the reduction in inflationary expectations so hoped for by the Council.

The greater reliance on fiscal measures may in part be due to the Council's disillusionment with monetary actions.

... the rate of expansion of the money supply [January-July 1971] had been extraordinarily large, much larger than had been commonly expected at the beginning of the year. Even when allowance was made for some lag between the increase in the stock of money and any consequent increase in economic activity, the response of the economy to the monetary expansion was less than many studies had found in previous experience. [pp. 65-66]

Whether judged by money supply or interest rate indicators, monetary policy has been highly erratic. The growth rate in  $M_1$  from January-July 1971 averaged about 11 percent while the last five months of the year saw the money supply virtually unchanged. The figures for the first quarter of 1972 suggest that the money supply has resumed a rate of growth approaching 9 percent. Short-term interest rates fell during the first quarter, then jumped

dramatically reaching a high point in July and fell back toward their lows by early 1972. The erratic swings in monetary indicators are unlikely to create a climate for stable expectations. The monetarists will argue that as long as the Federal Reserve is wedded to an interest rate target, stable growth of the money supply is impossible. The opponents will argue that money supply control is neither possible nor desirable and that interest rates and credit market conditions are the important factors in assessing monetary consequences. The Council's report leaves the impression that the Federal Reserve hoped to satisfy both schools by having their stated policy "... place emphasis on the monetary growth rate ..." while "... actual operations were designed to influence interest rates and conditions in short-term money markets, with the intention of thereby achieving the desired monetary growth rate" (p. 57). The ambivalences inherent in the Council's discussion of monetary policy (pp. 56-59) clearly reflect the lack of consensus among monetary economists. The Council finally settles for a "precept of steadiness with respect to monetary policy," and goes on to admit that:

The problem is that there is no single measure or objective combination of measures of monetary policy that is a completely satisfactory or completely superior measure of monetary policy by which a principle of steadiness could be calibrated. Judgement must be exercised. However, there is probably a presumption against extreme values or variations of the rate of change of narrowly defined money, i.e., currency plus demand deposits. [p. 112]

While the average rate of growth in  $M_1$  for 1971 of over 6 percent seems consistent with the Council's desire for an expansionary monetary policy, the shorter term fluctuations in the money supply clearly violate its desire for steadiness.

What then are the likely consequences

of the full package of policies introduced since August 15? The Council argues that:

Many aspects of the full package—that is, the freeze, the international measures, the fiscal steps—were unprecedented, and therefore any reliable calculation of the size of their effects was exceedingly difficult. The Council of Economic Advisers estimated at the end of August that the Administration's package would make real 1972 GNP in 1972 prices about \$15 billion greater than it would have been without the new program, but this figure was recognized as subject to a wide range of uncertainty. [p. 71]

Unfortunately, the Council does not reveal how this calculation was made. Had it done so, an evaluation of both its model and its forecast would have been much easier; but, as stated, the Council's assertion is totally untestable.

Given the Council's poor forecast record for 1971, one would have hoped for a more explicit discussion of its 1972 forecast, but there is not a single table presenting consistent forecasts for the components of GNP, nor is there a clear discussion of the component forecasts. Some forecasts are given for year-end figures, others for yearly averages, and still others for the fiscal year. Magnitudes are haphazardly reported in dollars, percentages, or percentage changes. It would therefore be most useful if future reports could include a technical appendix which presents a consistent set of component forecasts along with a description of the explicit model which generated those forecasts. Short of this, one must retreat to a more general evaluation of the consequences of the program.

Whether by accident or intent, the NEP package mirrors the insights gleaned from the theoretical literature on optimal policy decisions under uncertainty. Uncertainty calls for the simultaneous use of multiple policy instruments even when there is only

a single policy target (see William Brainard). The *NEP*'s forte in this context is its shotgun approach which dramatically utilizes many policy instruments to hit its multiple targets. The Council acknowledges the considerable uncertainties surrounding its marksmanship, but it never squarely confronts the question of which target is to be given priority. In the choice between inflation and unemployment, we can find little solace in the theoretical abstraction that in the "long run," any rate of inflation is consistent with "the natural rate of unemployment" (see Friedman). Given normal policy planning horizons, and the dynamic disequilibria that seem to characterize our economy, a trade-off between inflation and unemployment does exist, and the problem of establishing priorities cannot be finessed by relying on expectation effects which are presumed to improve the tradeoff. One would like to take the Council at its word when it expresses its intention that:

Reduction of the unemployment rate in 1972 is a primary objective of this year's economic policy. It is to this end that the Government is pursuing a highly expansive fiscal policy. And it is in large part to this end that prices and wages are controlled, so that the expansion of demand will generate more jobs, not more inflation. [p. 108]

But the Council still seems overridingly concerned with inflation and this has led it to overestimate the social costs of inflation:

The avoidance of inflation is always, of course, an objective of national policy, . . . . . But this objective may not get its proper weight because of failure to foresee the losses of output and employment that will later be entailed in ending the inflation. [p. 111]

The statement erroneously suggests that the costs of inflation must include the costs of future unemployment. But this result is

predicated on the false proposition that the inflation must be ended. Alternatively, if one begins with a menu of inflation rates and unemployment rates which are consistent over a reasonable policy planning period, the choice becomes one of the direct costs of inflation and the direct but separable costs of unemployment. Since it is most likely that the costs of sustained unemployment greatly exceed the costs of sustained inflation, the question becomes not how do we end inflation, but rather how can we learn to live with it (see James Tobin).

The Council's preoccupation with inflation has led it totally to ignore "poverty." The concept is absent from the report but the phenomenon is very much in evidence in the American economy. By ignoring the distributional consequences of unemployment, the Council fails to place the unemployment problem in a relevant social perspective. Adherence to a full employment target will undoubtedly involve inflationary pressures; however, it would be wiser to explore means of minimizing the costs of inflation than to live with the delusion that we can have both full employment and stable prices. Escalator clauses and interest payments on money can go a long way in reducing the distributional consequences of inflation. In contrast economists have found few remedies for the debilitating effects of sustained unemployment, and there are no ready means for recapturing the consequent losses of real output.

The Administration's actions in the past year are probably sufficiently expansive to move the economy in the direction of reducing unemployment, and given the excess capacity of the economy, further inflationary pressures are not likely to build up quickly. One can only hope that the Administration's priorities will continue to change so that sustained levels of higher unemployment will not become the

# The 1972 Report of the President's Council of Economic Advisers: International Aspects

By PETER B. KENEN\*

The international chapter of the Economic Report has rarely had more than a tenuous connection to the other chapters. Although the balance of payments has shown large deficits for many years, its impact on general policy has been hard to discern. In recent years, moreover, the usual statements of concern and pledges to restore equilibrium sounded even less forceful than usual, partly because the domestic dilemma preempted attention and partly because the balance of payments seemed to be improving. The balance on official reserve transactions showed substantial surpluses in 1968 and 1969 (see Table 1), and though there was a deficit in 1970, the deterioration did not evoke great concern; it was readily ascribed to short-term capital movements, and the trade accounts were getting better.<sup>1</sup> Indeed, one year ago, the editor of this *Review* saw fit to commission two papers on the Council's treatment of inflation and recession,<sup>2</sup> but none on the balance of payments.

This year was different. The entire Report is focused on the New Economic Policy announced in August of last year—

\* Professor of economics and international finance, Princeton University.

<sup>1</sup> None of the balances in Table 1, save perhaps the "basic balance," moved in a steadily ominous way to signal the crisis of 1971. And anyone who bothered to compute the basic balance before its recent debut in the official tables could have been excused for discounting its significance. It was wildly unstable in the first half of the decade. The Council puts it wryly: "Although the basic balance generally reflects underlying forces, it is sometimes subject to short-run movements" (p. 154).

<sup>2</sup> See Martin J. Bailey and Robert Eisner.

TABLE 1—SUMMARY OF THE U.S. BALANCE OF PAYMENTS, 1960–71  
(millions of dollars)

Year	Surplus or Deficit (–) on			
	Merchandise Trade	Basic Balance <sup>a</sup>	Net Liquidity Balance	Official Reserve Transactions
1960	4,906	–1,155	– 3,665	– 3,403
1961	5,588	20	– 2,229	– 1,348
1962	4,561	– 979	– 2,845	– 2,650
1963	5,241	–1,262	– 2,571	– 1,934
1964	6,831	28	– 2,745	– 1,534
1965	4,942	–1,814	– 2,493	– 1,289
1966	3,927	–1,614	– 2,148	219
1967	3,859	–3,196	– 4,685	– 3,418
1968	624	–1,349	– 1,610	1,641
1969	660	–2,879	– 6,084	2,702
1970	2,110	–3,038	– 3,821 <sup>b</sup>	– 9,821 <sup>b</sup>
1971	–2,879	–9,284	–21,973 <sup>b</sup>	–29,767 <sup>b</sup>

Source: *Survey of Current Business*, June 1971 and March 1972.

<sup>a</sup> Balance on current account and long-term capital.

<sup>b</sup> Allocations of Special Drawing Rights treated as capital inflows, understating deficits (by \$867 million in 1970 and \$717 million in 1971).

on its rationale, content, and probable consequences—and the international financial situation is vital throughout. International considerations appear in every chapter, even the one on resource allocation (ch. 4) which stresses connections between the competitive position of American industry and its spending for research and development. In more than one place, indeed, the Council itself suggests that the balance of payments governed the timing of the shift in policy:

More crucial than either of these [inflation or unemployment] for the timing of the decisions was the serious weakening of the dollar in international markets. . . . Funds totaling about \$3.7 billion moved into foreign official reserve accounts in the week ended August 15. The time had come to deal decisively with the international financial problem that had persisted for at least a dozen years despite the efforts of four successive Administrations. [pp. 29-30]

One is even led to wonder whether this Administration could have introduced its new domestic policies had the foreign-exchange markets not supplied a compelling excuse to set aside its earlier predictions and commitments. The Council is at pains to contradict this supposition:

The suspension of dollar convertibility and the freeze were the dramatic elements in the announcement, and this might have led to an impression that the program was aimed primarily at solving the problems of the balance of payments and inflation rather than the problem of unemployment. In fact, the program was directed at all three problems. [pp. 22-24]

Yet the sudden and massive deterioration of the balance of payments must have strengthened greatly the influence of those in the Administration who in the interest of domestic expansion (a) were willing to abandon convertibility, (b) had favored for some time and sometimes in public a shift from monetary to fiscal policy, (c) had favored an incomes policy to slow inflation during the recovery.

The international chapter of this year's Report deals sequentially with four major subjects: (i) the immediate and long-term causes of the international financial crisis; (ii) the debate and negotiations concerning a solution, culminating in the Washington agreement of December 18 on new exchange rates; (iii) the likely consequences of that agreement; and (iv) the continuing need for reform of the monetary system. Each is treated carefully, even candidly,

but there are several serious sins of omission, and some of these appear to be committed deliberately.

In its account of long-term trends leading to the payments crisis, the Council gives greatest weight to the deterioration of the trade balance and argues persuasively that cost and price developments take most of the blame. It is difficult to make the link precisely, despite the spate of recent econometric work on trade and payments. Comparisons with the early 1960's, when the trade balance was strong, are impaired by a cyclical configuration different but no less extreme than the current situation (and the trade balance then was markedly stronger than it had been in the late 1950's). The data for 1971 are badly distorted by direct and indirect effects of the dock strikes. Finally, the indexes of export unit values used to measure the competitive effects of inflation (p. 152) are notoriously unreliable.<sup>3</sup> Yet the change in this price index is too large to be ignored; it rose by 26.7 percent from 1964 through the first half of 1971, compared to 17.7 percent for other industrial countries.

The Report, however, may explain too much, at the expense of other developments deserving attention in the diagnosis and in the subsequent appraisal of devaluation. Thus, the increase of automotive imports accounted for a third of the growth in total imports during 1971 (despite a sharp fall in the fourth quarter, following repeal of the excise tax).<sup>4</sup> In fact, the behavior of automotive imports was similar to that of the late 1950's, during the earlier deterioration of the trade balance and prior to the advent of the first domestic compacts. More importantly, the Report does not furnish the trade data required to

<sup>3</sup> See, e.g., Irving B. Kravis and Robert E. Lipsey, pp. 4-6.

<sup>4</sup> See *Survey of Current Business*, Mar. 1972, pp. 38-40.

appraise three of its principal assertions—that the erosion of the current account was due to inflation, that cyclical and other corrections to the current account revealed a “normalized” deficit of \$0.6 billion in 1970 and projected a normalized deficit of \$4 billion in 1972 (pp. 152–54), and that the change in exchange rates posted in December cannot fully achieve the \$13 billion turnaround proclaimed as the aim of American policy (p. 161).

The Council attributes the crisis itself to a converging consensus among “individuals, firms and governments” that “the value of the U.S. dollar was going to fall relative to the other major currencies” (p. 144), and traces the consensus to three distinct opinions—that “. . . the United States had been in fundamental disequilibrium throughout the 1960’s” . . . that “. . . the poor wage-price-productivity performance of the U.S. economy between 1965 and 1969 . . . had significantly lowered the competitiveness of U.S. goods . . .” and that “. . . developments in the conduct of monetary policy here and abroad . . . would induce large outflows of short-term capital from the United States to Europe” (pp. 144–45). Strategic to this survey, however, is not the list of reasons, but the list of participants. If central banks and governments had been prepared to accumulate dollars without stirring restlessly, the opinions of individuals and firms could not have forced devaluation, even if they were well founded. The behavior of governments, moreover, was not necessarily due to a change in their opinions. A change in their circumstances, mentioned quite casually (p. 146), was quite sufficient. During 1970, their dollar holdings rose by \$7.3 billion, more than in either of the previous half decades taken as a whole.

When, five years ago, the Federal Reserve began again to slow the growth of the credit base, *U.S.* commercial banks re-

sponded by borrowing from the Euro-dollar market, directly and through foreign branches. Their obligations to their branches rose from \$4.2 billion in December 1967 to \$14.3 billion in September 1969. At that point, further borrowing was made subject to reserve requirements, and the process came to a halt, but the banks’ indebtedness to their branches remained above \$12 billion throughout the first half of 1970. Many of the borrowed dollars, moreover, came indirectly from foreign official holdings, as private foreigners bought dollars to deposit in the Eurodollar market. Liquid liabilities to foreign official institutions fell from an historic high of \$15.6 billion in December 1967 to a low of \$10.2 billion in June 1969. European reserves fell quite sharply and, more to the point, their dollar holdings dropped as a fraction of reserves. When the Federal Reserve reversed direction in 1970, commercial banks reduced their Eurodollar debts; obligations to branches fell by \$4.5 billion in the second half of 1970 and by another \$6.3 billion in the first half of 1971. Combined with other outflows, some of them related to the same change in policy, these repayments added hugely to the reserves and dollar holdings of foreign official institutions. The latter reached \$20.1 billion in December 1970, and climbed to \$30.6 billion in June 1971. (This despite the sale of \$3 billion in Treasury and Export-Import Bank securities to the foreign branches of *U.S.* commercial banks.) Gold conversions were not large in 1970, for dollar holdings were not yet abnormally high in relation to reserves, but began to grow in size and number in 1971. Central banks and governments were nervous, and conveyed their doubts to the foreign-exchange markets.<sup>5</sup>

<sup>5</sup> The fact that our gold stock was nearing \$10 billion, long viewed as a magical number in some circles abroad, did not help to reduce anxiety in private and official quarters.

The Council says little about the mix of policies that played so large a role in this important episode. It mentions German efforts to combat inflation by maintaining high interest rates in the face of a capital inflow, and the strategic decision to unpeg the mark which followed in May and introduced the interval of massive speculation (p. 147). But other countries' errors, even when repeated, must be treated politely; the Council refrains from reminding us that the Germans had done this before, with the same unsettling consequences. Yet it fails also to recall its own earlier criticism of American policy—the statistical miscalculation of monetary growth in the winter of 1970–71 which, it alleges, led to excessive expansion in the spring (p. 57); this expansion worked to perpetuate the large differences in short-term interest rates between the United States and other countries and was therefore responsible for some of the capital outflow that set the stage for speculation. The changes in exchange rates of May and December were overdue, but the authorities can take little credit for them. They were backed into progress by a persistent, indefensible reliance on monetary policy to accomplish their domestic aims. Those who despair of converting governments to the virtues of flexible exchange rates are perhaps less deserving of sympathy than those who have tried and failed to make a modest case for fiscal policy and for the optimum policy mix in open economies.

The Report's description of events from August to December does not evoke a single recollection of our bad manners—the unilateral abrogation of explicit obligations to the International Monetary Fund and conduct damaging to allies whose policies and politics require continuing evidence of our consultation and cooperation. Its blandness, however, seems almost to mock the strident language heard at the time, when the President and Mr. Con-

nally charged repeatedly that *U.S.* exports were treated unfairly and, early on, demanded unilateral commercial concessions. The Council quotes Mr. Conally:

If other governments will make tangible progress toward dismantling specific barriers to trade over coming weeks and will be prepared to allow market realities freely to determine exchange rates for their currencies for a transitional period, we, for our part, would be prepared to remove the surcharge. [p. 159]

It invites us, however, to infer dissatisfaction with this particular tactic. Thus in its concluding section on trade policy, we hear a stifled sigh of relief:

For over three decades the free world has been gradually liberalizing commercial policy. . . . . This course recently has faced and survived two major tests. There have been strong pressures in the United States itself to redirect policy toward quantitative restrictions on trade and investment. While the United States extended import restraints to a few additional commodities in 1971, legislation that would apply quantitative restrictions broadly to imports has not been enacted. The second major test . . . came with the trade policy actions to combat the *U.S.* payments deficit taken in connection with the New Economic Policy. Although it provoked considerable controversy, the import surcharge did not set off the series of retaliatory actions abroad that had been feared in some quarters. And it was removed before the end of 1971 when the currency realignment made it no longer necessary. [p. 164]

A long passage on trade in farm products (pp. 170–72) does describe the ways in which price-support policies distort foreign trade, but it is equally critical of European and American measures (the latter for restricting exports, as well as imports) and closes by seconding the even-handed judgments of the President's Commission on International Trade and Investment Policy

(1971), not the one-sided views heard elsewhere last year.<sup>6</sup>

Turning next to the aftermath of August, the Council deals at length with the realignment. The debate on exchange rates, it says, focused on three questions—the choice between floating rates and negotiations as ways to reach new parities, the choice between a revaluation of other currencies and a devaluation of the dollar, and the size of the requisite change in exchange rates, by whatever way achieved. Then it lists reasons, given or implied, for choosing negotiations and devaluation, stressing those considerations and constraints which academic advocates of floating rates deprecate or ignore (to the detriment of the

<sup>6</sup> Unhappily, the rest of the section on foreign trade is less comforting to advocates of liberal trade policies. It offers a pathetic defense of the Domestic International Sales Corporation created to shelter export profits from U.S. income tax: Because the profits of overseas subsidiaries are not subject to U.S. tax until they are repatriated, domestic manufacturers selling abroad in competition with them are entitled to similar tax deferral (pp. 167–68). Tax neutrality between the domestic and foreign production of American firms is, of course, an appropriate aim. It is chiefly important, however, to eliminate the tax inducement to overseas investment, not to promote exports, and should not be achieved by upsetting neutrality in the tax treatment of home and export sales by domestic firms. Why not remove the tax advantage presently enjoyed by the foreign subsidiaries, as proposed by the Kennedy Administration in 1961, or are we always doomed to correct old inequities by creating new ones?

Later, we encounter a subsection on “easing the adjustment to imports” and are invited momentarily to expect proposals for the liberalization of adjustment assistance or, at least, a promise of new studies responsive to objections raised by and before the President’s Commission. Instead, we are treated to a discussion, studiously balanced, of “voluntary” export quotas like those enshrined in the Long-Term Agreement on Cotton Textiles and newly extended to man-made and wool products. The Council is understandably concerned to ward off pressures from Congress and others for comprehensive import restrictions, the same concern that must have figured prominently in the efforts of the President and Secretary Connally to obtain trade concessions before lifting the surcharge, and it argues quite properly that the realignment will itself diminish the need to protect American industry from “market disruption” (p. 170). But it is excessively defensive and stresses third-best ways to correct distortions and to redress injury.

academics’ relevance and influence). Thus, floating rates for the transition were much mistrusted, especially by those who were already floating, as the French and others had restricted capital inflows and partitioned the foreign-exchange markets to hold down their own rates. In consequence, countries with floating rates and others that might have considered them faced larger appreciations against the dollar and third currencies (especially the franc) than were wanted or needed over the long run. The Report goes on to remind us that the case for a devaluation of the dollar was mainly political: “. . . if the United States devalued, other countries could reduce or avoid the political onus of revaluing” and the dollar would no longer appear to be “. . . the benchmark against which all other national currencies set their values” (p. 160). These arguments have little to commend them, but the Council is right to stress their pervasive appeal and stubborn power. Thereafter, however, it goes too far, echoing an argument that should have expired when the central banks ceased to be private institutions:

. . . revaluation reduces the value, expressed in the domestic currency unit, of a nation’s stock of foreign monetary assets. . . . This balance-sheet loss for each revaluing nation, so far as the gold component is affected, could be reduced by requiring the United States itself to contribute to a realignment through an increase in the dollar price of gold. [p. 160]

Should central banks be worried about their net worth? Or are we being asked to worry instead about the net worth of others who hold gold?<sup>7</sup>

<sup>7</sup> The Report makes no mention of other arguments, slightly more sensible. If the dollar had not been devalued, it would have been necessary to revalue many foreign currencies, including those of countries outside the Group of Ten, and this would have required negotiations with numerous governments. The values of the currencies in question have, of course, changed with the devaluation of the dollar, and this has required the tacit

Finally, the Council tries to show us how the United States came to demand a \$13 billion improvement in its balance of payments, but generates more puzzlement in the process. In the absence of cyclical and other special circumstances, there would have been a \$4 billion current-account deficit in 1971. But we needed a \$9 billion surplus. Some \$2 billion was required to cover persistent short-term capital flows, and \$1 billion was required to offset the average adverse balance on net errors and omissions. The largest part, however, was required to finance our aid and long-term lending to the less developed countries:

consent of the same governments, but it has not required initiatives. Some, of course, have taken action on their own, declaring new par values or central rates in order to alter the dollar prices of their currencies by more or less than 8.57 percent, the amount by which the dollar was devalued in terms of gold, but this was a piecemeal process, simpler than negotiation. As of January 17, 1972, the majority of countries outside the Group of Ten had devalued their currencies by as much as the dollar or left their parities unchanged (and these two groups were very nearly equal in numbers); only eight had devalued by more than the United States, eight others had devalued by less, and two had revalued.

There are, in addition, "balance-sheet" arguments better than the one mentioned by the Council. Devaluation in terms of gold raised slightly the dollar value of U.S. reserve assets relative to U.S. liabilities. More importantly, it served to maintain the "average" purchasing power of reserve claims on the International Monetary Fund (IMF). In the words of Pierre-Paul Schweitzer, Managing Director of the Fund:

It would be desirable . . . if all the major countries involved were to make a contribution to the realignment of currencies, not only for reasons of equity, but also for achievement of an appropriate relationship of currencies to gold and, what is perhaps more important, the SDR's and reserve positions in the Fund. Moreover, if one contemplates an enlarged future role for SDRs in the international monetary system, they must continue to be attractive reserve assets—which would require that they hold their value against currencies in general.

The Council hints at this point in the passage on the benchmark quoted above. "There was a body of opinion in Europe that the benchmark should be an objective one, or at least a multinational one . . . . . The existence of this body of opinion has important implications for the choice of a basic monetary unit of account in the international system of the future" (p. 160). But it does not mention SDR's in this context (and, more disturbing, mentions them only twice in the entire section on reform).

. . . The annual outflow for Government grants and credits plus private long-term capital flows from the United States to countries other than Western European nations, Canada, and Japan was estimated at \$6 billion, or just over one-half of 1 percent of the U.S. gross national product. The average annual outflow for these purposes during the 5-year period from 1967 through 1971 was about \$5½ billion. [p. 154]

But do we really grant, lend, and invest more than \$5 billion in the less developed countries? And how can we continue to invest in the developed countries if our entire current-account surplus is to be earmarked for other purposes? Surprisingly, we do and can—or did and could in recent years. In 1970, for example, the *gross* outflow of long-term funds to the less developed countries was \$6.5 billion, and the *net* outflow was \$5.2 billion (see Table 2). But repayments of past aid *plus* long-term foreign lending and investment in the United States covered U.S. lending and investment in Europe, Canada, and Japan; there was no net outflow. The data given by the Council are not wrong nor misleading. If they are less than enlightening, it is because they are constructed to persuade, rather than inform. The United States does not dare ask Europe to help us become more competitive in order that Americans can "buy up" Europe's industry. It can ask help, however, in the name of the less developed countries.<sup>8</sup>

<sup>8</sup> One is still puzzled, however, by the Council's assertion that \$2 billion is required to finance persistent outflows of U.S. short-term capital. From 1964 through 1970, these outflows averaged a bit less than \$1 billion, and were much more than offset by inflows of foreign short-term capital. The annual averages (in millions of dollars) were:

Increase of U.S. short-term claims (lines 42-43, 45-46) . . . . .	826
Increase of foreign nonliquid claims (line 51) . . . . .	391
Increase of foreign liquid claims (line 56) . . . . .	1,391

(Data and line numbers from *Survey of Current Business*, June 1971, Table 2.) If one could not anticipate any steady inflow of foreign capital, the United States would need only \$800 million to offset the gross outflow; if one

TABLE 2—LONG-TERM CAPITAL TRANSACTIONS IN THE  
U.S. BALANCE OF PAYMENTS, 1970  
(millions of dollars)

Item	To (+) or from (-)	
	Less Developed Countries <sup>a</sup>	Developed Countries
Gross Outflow of U.S. Capital	6,462	4,370
Government grants, except military	1,700	39
Government credits, including short-term (net)	2,738	574
Private direct investment <sup>b</sup>	1,494	2,951
Other private long-term capi- tal, net <sup>c</sup>	530	806
Repayment of U.S. Government Credits	- 883	- 836
Gross Inflow of Foreign Capital	- 359	- 3,518
Private direct investment <sup>b</sup>	- 42	- 927
Purchases of U.S. securities, except Treasury issues	- 500	- 1,690
Other, net <sup>d</sup>	183	901
Net Outflow (+) from the United States	5,220	16

Source: *Survey of Current Business*, March 1972.

<sup>a</sup> Includes Latin America, Australia, New Zealand and South Africa, other countries in Asia and Africa (except Japan), international organizations and unallocated.

<sup>b</sup> Excludes the reinvested earnings of subsidiaries.

<sup>c</sup> Includes amounts on lines 40, 41, and 44 of *Survey* Tables 2 and 9.

<sup>d</sup> Includes amounts on lines 50 and 55 of *Survey* Tables 2 and 9; amounts on line 52 (long-term liabilities to private foreigners reported by U.S. banks) should be included, but cannot be separated by region.

Although we are told that the realignment of exchange rates will not bring in \$13 billion, the Council does not tell us how much to expect. The Report projects no net change in the current-account balance; last year's balance was exactly zero, and the average for 1972, it says, "... will

could anticipate an increase of nonliquid foreign claims, but no further build-up of liquid claims (including claims by the U.S. banks' branches abroad), it would need only \$400 million. Does the \$13 billion target, including the \$2 billion, allow for an increase of short-term outflows or, perhaps, the amortization of foreign dollar holdings (including official reserve holdings)?

probably be close to zero" (p. 104). We are warned, however, that "The expected adjustment of U.S. net exports might involve a longer lag than we have posited" (p. 107). This would mean a deficit for 1972.

One can perhaps infer an expectation as to the swing in the current account over the course of the year. The Council's forecast for the year as a whole would require a \$7 or \$8 billion improvement during the year, from a deficit of \$4.6 billion per annum in the fourth quarter of 1971 to a surplus of \$2 or \$3 billion in the fourth quarter of 1972.<sup>9</sup> Any smaller surplus at the end of the year would not forestall a year-long deficit. But the Council does not help us to decide how much of this swing can be ascribed to the realignment (and to the trade-policy changes still being negotiated) and how much more can be expected in 1973.<sup>10</sup>

The Council devotes more space to the prospects for reform of the monetary system than to the prospects for the balance of payments. Here, however, it sins again by omission. Its agenda for reform stresses quite properly the question of exchange rates—when and how to change them. But it gives inadequate attention to the problems of liquidity, convertibility, and consolidation. It does devote a paragraph to the need for agreement on the level and composition of reserves, and refers obliquely to the chief issue:

The United States and many other countries share the conviction that gold should and will play a diminishing role in the system. Already, considerable progress has been made in developing

<sup>9</sup> This would not be large, even by recent standards; there was a \$2.7 billion surplus in the fourth quarter of 1970, and a \$3.6 billion surplus for the year.

<sup>10</sup> If, of course, the Council is correct in warning that cyclical developments could be damaging in 1972, the \$7 to \$8 billion improvement mentioned above would represent a larger gain on a cyclically adjusted basis, amounting to much more than half of the \$13 billion turnaround sought in the negotiations.

the SDR as an alternative international reserve asset, but many questions remain, including the appropriate role of the dollar and other reserve currencies.  
[p. 164]

But we come away from the Report too close to innocence. Could any reader know that the role of dollar, not gold or *SDR's*, is the critical, divisive problem? The Council, like the Treasury, seems no less obdurate in refusing to consider the question of convertibility than it was before August in refusing to consider the question of suspension.

No one able to add or subtract could ask that the United States reopen the gold window. At the end of 1972, *U.S.* liquid liabilities to foreign official institutions totalled \$47.0 billion. Its total reserves were just \$12.2 billion, of which only \$10.2 billion was gold.<sup>11</sup> But something must be done. Countries holding dollars cannot use them freely, even for transactions with the *IMF*. And countries are obliged to acquire more dollars, without the prospect of quick use or conversion, whenever they must intervene in the foreign-exchange markets to preserve the exchange rates established a few months ago.<sup>12</sup>

<sup>11</sup> These figures do not reflect the revaluation of our gold stock resulting from the legislation raising the gold price, but this gain was small (and was offset in part by larger liabilities to the *IMF*, resulting from the same revaluation).

<sup>12</sup> They did acquire dollars in the early months of 1972 when, as in 1971, interest rates and rumors combined to cause a temporary flight from the dollar. During that same episode, however, the dollar was allowed to fall very rapidly through the new wider permissible range. One wonders if foreign central banks were reluctant to intervene until they were obliged to do so, because they did not want to buy more dollars (and whether their reluctance to intervene caused the flight from the dollar to snowball briefly). Note, finally, that the Council's judgment concerning the (inadequate) size of the realignment is an implicit forecast of additional dollar deficits and an additional reason to doubt the quality of the dollar as a reserve asset. Is this perhaps the reason for refusing to discuss convertibility; does the Administration hope to finance further deficits by forcing inconvertible dollars into the coffers of foreign central banks unwilling or unable to contemplate another round of

This is not the place to survey all the ways in which the problem might be solved. It is, perhaps, the time to make a complaint. The United States has lost the initiative. When it closed the gold window and called for realignment, most observers expected that the future of exchange rates and of the dollar would be settled together. There was at least an expectation of agreement in principle on steps toward long-run reform, to be announced along with the realignment. Just one month before the realignment itself, Frank Southard, Deputy Managing Director of the *IMF*, described succinctly the prevailing view:

[S]ome understanding will need to be reached with respect to the nature and extent of convertibility of official dollar holdings, both as to those existing at the time of any realignment and as to dollar accruals thereafter.

As for the longer-term stage, involving more basic reforms in the international monetary system . . . [it] may be reasonable to assume that neither the United States nor any other country will again freely buy and sell gold as the sole means of maintaining convertibility of currencies. But some form of effective convertibility will certainly be needed and this relates especially to the role of reserve currencies and particularly to the role of the *U.S.* dollar in the future.

Specific suggestions came from many quarters, but had in common several innovations that would have thrust *SDR's* toward the center of the monetary system, displacing gradually gold and national currencies and providing the only orderly way of enlarging global reserves—including and especially *U.S.* reserves. The

---

revaluations? Or is it waiting merely for the long-awaited reflux—the reconstitution of private dollar holdings—to reduce foreign official holdings and minimize the size of the external debt it must consolidate? If so, it may be disappointed, for the reflux may itself await governmental agreement on the future of the dollar in the monetary system.

Chancellor of the Exchequer, Anthony Barber, spoke for many when he presented his own plan to the 1971 Meetings of the International Monetary Fund:

First, the SDR could become the numeraire in terms of which parities are expressed and in relation to which currencies are revalued or devalued.

The second point is that the SDR could become the main asset in which countries hold their reserves . . . with currency holdings largely confined to working balances.

Thirdly, arrangements would be needed to provide for the controlled creation of adequate but not excessive world liquidity. . . . This was indeed the intention of the existing SDR scheme. . . .

The United States, however, held back from endorsing any long-run reform, and even discouraged discussions in the Group of Ten which, it said, was unrepresentative. Instead, the Administration used its enormous bargaining power to press for trade concessions, along with revaluation.

This position is not hard to understand. Until one can be sure that the realignment and other measures have corrected the fundamental disequilibrium in the U.S. balance of payments, the United States may stand to gain, not lose, by delaying reforms which could make it more difficult to finance U.S. deficits. If, further, the balance of payments begins soon to improve, causing a "reflux" of dollars, there will be fewer dollars in foreign banks, and the problem of consolidation will be diminished in size and difficulty.

Yet delay is costly too, for the atmosphere is changing. The spokesmen for gold have come forward again, some to demand the implausible—a return to unlimited convertibility at the new gold price. The enthusiasm for SDR's seems to have waned, and, with it, the prospect of enlarging U.S. reserves. The negotiations on reform will be more difficult than they might have been, and the mood less con-

ducive to the radical reforms that seemed possible less than a year ago.

The negotiations cannot be delayed much longer—even though the atmosphere is much less propitious. The realignment itself is not likely to survive if foreign central banks are required to support it by purchases of dollars they can neither use nor transform into assets they want to hold. Worse yet, it may survive formally, but only because it is deprived of its chief significance; there is a very real risk of widespread resort to exchange controls and the further fragmentation of the foreign-exchange markets, as governments and central banks strive to avoid the need to accumulate additional dollars.

One can only hope that these warnings will be rendered obsolete before they are published—that the Administration will come to fear the risk it is running. The inconvertible dollar of 1972 is not a better basis for the monetary system than the convertible dollar of earlier years. It is, indeed, inferior politically and financially.

The introduction of limited flexibility into the exchange-rate system—the aim to which the Council devotes most attention—cannot substitute for measures to enlarge U.S. reserves, to restore the transferability or convertibility of newly acquired dollars, and to deal with the overhang of dollar balances created by past deficits, thereby to free U.S. reserves for the financing of future imbalances.

#### REFERENCES

- M. J. Bailey, "The 1971 Report of the President's Council of Economic Advisers: Inflation and Recession," *Amer. Econ. Rev.*, Sept. 1971, 61, 517-21.
- A. Barber, "Statement to the Annual Meeting of the International Bank for Reconstruction and Development and the International Monetary Fund," IBRD-IMF Press Release No. 14. Sept. 28, 1971.
- R. Eisner, "The 1971 Report of the Presi-

- dent's Council of Economic Advisers: Inflation and Recession," *Amer. Econ. Rev.*, Sept. 1971, 61, 523-26.
- I. B. Kravis and R. E. Lipsey, *Price Competitiveness in World Trade*, New York 1971.
- P.-P. Schweitzer, "Statement to the Annual Meeting of the International Bank for Reconstruction and Development and the International Monetary Fund," IBRD-IMF Press Release No. 2. Sept. 27, 1971.
- F. A. Southard, Jr., "Remarks at the National Foreign Trade Convention," in International Monetary Fund, *International Financial News Survey*, Nov. 24, 1971.
- Commission on International Trade and Investment Policy, *United States International Economic Policy in an Interdependent World*, Washington 1971.
- U.S. Council of Economic Advisers, *Economic Report of the President*, Washington, Jan. 1972.
- U.S. Department of Commerce, *Survey of Current Business*, various issues, Washington 1971-72.

# The 1972 Report of the President's Council of Economic Advisers: Inflation and Controls

By REUBEN A. KESSEL\*

If there is a theme to the Economic Report of the President's Council of Economic Advisers, it is—we are trying to and succeeding in breaking inflationary expectations. The foreword by President Nixon says: "We are converting the fear of perpetual inflation into a growing hope for price stability" (p. 5). This emphasis on eliminating expectations of inflation occurs again and again in the Report. To illustrate:

From mid-August to the end of the year, there was slow but steady improvement in confidence that the rate of inflation was subsiding and the pace of the economic recovery was gathering strength. [p. 29]

Probably the most important justification for the price freeze that appears in the Report deals with expectations. The justification for the freeze is:

... to reestablish an acceptably low rate of price increase by reducing expectations of continued strong inflation and eliminating, to the extent possible, practices and behavior which sustain or promote inflation. [p. 83]

The basic premise of the price-wage control system is that the inflation of 1970 and 1971 was the result of expectations, contracts, and patterns of behavior built up during the earlier period, beginning in 1965, when there was an inflationary excess of demand. . . . . The purpose of the control system is to give the country a period of

enforced stability in which expectations, contracts, and behavior will become adapted to the fact that rapid inflation is no longer the prospective condition of American life. [p. 108]

Whether the authors of the Economic Report of the President believe price and wage controls can stop inflation as distinguished from inflationary expectations is unclear. The passage just cited suggests that direct controls can prevent inflation. However, other passages dealing with the same question are more ambiguous. Consider:

If monetary and fiscal policy keep the growth of demand moderate, the price and wage controls can bring about more quickly and surely the lower rate of inflation that competitive forces would cause in such circumstances. [p. 96]

However,

... if demand is allowed to grow excessively, the price and wage control system will lose its value. [p. 96]

And,

... the existence of price and wage controls will reduce the pressure both of inflation and of inflationary expectations. [p. 101]

There is no hedging by the authors of the Report in their overall evaluation of the merits of the freeze. "In the accomplishment of its own objectives [presumably the objectives of the Administration], the freeze was an unqualified success" (p. 96). "Viewed as a whole and against the

\* Professor of business economics, University of Chicago.

circumstances of its initiation, the freeze was a major success" (p. 81). And, "... the freeze program must be accounted outstandingly effective" (p. 80). Evidence to support these glowing evaluations was not presented.

The Report also has some of the simplicity of an old western. The villains, figuratively attired in black hats, caused "... price and wage increases that were not justified by competitive market conditions and were helping to prolong the inflation and unemployment" (p. 21). These villains can be prevented from causing inflation, cut off at the pass, through the use of "emergency expedients" which is a euphemism for direct controls (see p. 24). The heroes of the Economic Report are the Cost of Living Council, the Pay Board, the Price Commission, and above all, the "... self-restraint of the American people" (p. 100).

One can surmise that the authors of the Economic Report are intimating that greed is the source of inflation and if self-restraint were exercised in the pursuit of self-interest, controls would be redundant. Precisely how inflation is caused by price and wage increases that are unjustified by competitive market conditions is unspecified. How does monopoly pricing in wage and product markets cause inflation? One will search in vain for an answer to this question in the Report. If noncompetitive pricing implies inflation, then the fraction of the economy monopolized must have declined sharply following World War II and the Korean War and grown again in recent years. Some of the noncompetitive price increases that have occurred in recent years, according to the authors of the Report, are a result of governmental intervention in specific markets. An increase in the past six years of over 50 percent in price support for milk is cited (see p. 171). How this important source of noncompetitive pricing and price increases will be

restrained by direct controls is not discussed.

Since the theme of the Economic Report is that wage and price controls will break inflationary expectations, it is relevant to ask 1) to what extent do inflationary expectations exist, and 2) if they exist, will they be broken by the direct controls imposed? Ideas useful for answering both questions appear in chapter nineteen of Irving Fisher's *Theory of Interest* which sets forth Fisher's empirical findings and is to this day one of the best works available for understanding the formation of inflationary expectations.

Briefly the message in Fisher is, if inflationary expectations exist, then they can be detected in the money and capital markets. To the extent that inflation is anticipated, nominal or money rates of interest will exceed real rates by the expected rate of inflation (see p. 411 ff). Fisher's empirical findings indicate that current expectations of inflation are an average of past rates of change of prices (p. 438). Hence Fisher's results imply that if the rate of inflation is accelerating, as was the case in the five or six years before the imposition of controls, the market will underanticipate the rate of change of prices and debtors will expropriate creditors. The change to a higher rate of inflation implies, according to Fisher's results, that the yields observed in the money and capital markets contain downward biased estimates of the actual rate of inflation.

Evidence of the implications of Fisher's work for the accelerating rate of inflation of recent years can be obtained from our money and capital markets. For the six years from 1966 through 1971, one-year interest rates averaged about 5.7 percent.<sup>1</sup> The annual rate of change of prices during

<sup>1</sup> Council of Economic Advisers, p. 262, Table B-57, from the column on 9-12 month issue, and Table B-45, p. 247.

these years, as measured by the change in the Consumers Price Index (*CPI*), averaged about 4.5 percent.<sup>2</sup> Hence the real rate of interest, i.e., the rate of interest abstracting from inflation, was about 1 percent. Since this period was on the whole highly prosperous, it is unreasonable to argue that the economic value of one-year money was as low as 1 percent. Consequently, one must conclude that Fisher was correct when he argued that the market would underanticipate inflation when the inflation rate was accelerating.

The view that what occurs in money and capital markets is an indicator of the expected rate of inflation is not completely absent from the Report. What is absent is an interpretation of the relationship consistent with Fisher. Consider these contrasting statements:

After March, interest rates began rising as a result of the reversal of inflationary expectations. . . . .<sup>3</sup>

The decline of interest rates may have reflected some abatement of inflationary expectations. [p. 91]

Elsewhere the Report attributes correctly the fall in interest rates in 1970-71 to the recession (see p. 145). Since the end of World War II, rates have tended to be low about cyclical troughs and high about cyclical peaks. Hence, the fall in rates that occurred in 1970-71 is wholly attributable to the recession and is consistent with no change in expectations.

Fisher's long lags imply that it takes many years to either incorporate expectations of inflation into interest rates or remove them once they are incorporated.

<sup>2</sup> Some recent research suggests that the *CPI* may overmeasure price increases because of the way new quality improvements are incorporated into the index. However, the exclusion of governmental services from the *CPI* probably leads to an undermeasurement of the actual rise of prices. The net effect of these contrasting biases is unclear.

<sup>3</sup> See p. 60. To this writer, it was not clear from the context whether or not this was a typographical error.

Hence even if our experience with controls is interpreted by the market as equivalent to price level stability, the period of controls is too short to have anything but a negligible effect on rates.<sup>4</sup>

The secular trend upwards in interest rates since the end of World War II, particularly the rise of short relative to long rates, suggests that some expectations of inflation have been incorporated into the term structure of interest rates. Nevertheless, given the acceleration in the rate of inflation that occurred during the six years previously cited, the expectations incorporated in interest rates have been downward biased.<sup>5</sup>

The Report implicitly assumes that the market will regard direct controls as producing price level stability or reducing inflation and consequently lead to expectations of price level stability. However, if the Fisher view is correct, the market will look back to its previous encounters with direct controls in forming their expectations about how much inflation to expect. This implies that the experience of the market with direct controls during World War II and the Korean War is relevant.

These episodes, particularly World War II, suggest that price controls will camouflage but not prevent inflation. During World War II, it was the price indexes that were stabilized; what happened to prices in an economic sense was difficult to determine. After controls were removed and the *CPI* rose, the extent to which inflation was concealed became apparent.

<sup>4</sup> It is important to recognize, as Fisher did not, that the long rate of interest should reflect long-term expectations of inflation and the short rate, short-term expectations. Hence, if one is trying to determine what the market thinks the rate of inflation will be in the next year, then it is the one-year rate that it is relevant to examine.

<sup>5</sup> William Gibson shows that short rates incorporate expectations more quickly, i.e., have shorter lags, than long rates (p. 19).

However, on the surface it appeared as if inflation occurred when controls were removed and not during World War II. If Fisher is right and the market uses its past experience with price and wage controls in forming price level expectations, and if the objective of the current controls is to reduce inflationary expectations, then the current controls are counter-productive. Under these assumptions, their initiation by the government would be interpreted as an indicator of intentions to inflate and to disguise the inflation produced.

One could argue that the concern expressed in the Report with expectations of inflation reflects concern with the efficiency losses associated with anticipated inflation. Anticipated inflation leads to a substitution of scarce resources for money as a result of the rise of costs of holding money; it causes the private costs of holding money to rise relative to the economic or social costs. Hence economically scarce resources are substituted for relatively free resources. However, one will search in vain for a discussion of the efficiency losses caused by the anticipation of inflation.<sup>6</sup>

The Report seems to intimate that it is the expectations of inflation that have caused some of the inflation that has occurred in recent years. No evidence is presented to support this view. If the findings presented here are correct, expectations of inflation have not had a strong enough effect on interest rates to leave debtor-creditor relations undisturbed. Consequently, there have been redistributive effects that are difficult to defend.

The Report is almost totally silent on the causes of the inflation that this country has experienced in recent years.<sup>7</sup> On this

point, it is considerably less than candid. The inflation of the last half-decade was made in Washington and is the result of governmental monetary and fiscal policies and not monopoly pricing in product or factor markets, or expectations of inflation by the public. Insofar as expectations of inflation exist, they are, according to Fisher, a delayed response to inflation and not a cause of inflation.

The statements in the Report on the role of controls indicate that the authors of the Report regard price and wage controls and the new economic policy, of which it is a part, as providing a degree of freedom to pursue monetary and fiscal policies to restore full employment that could not be pursued in the absence of controls.<sup>8</sup> Hence, they either believe that controls are effective in preventing, as distinguished from disguising, inflation or they believe that the disguise will be effective; it will forestall public outcries about inflation. In either case, the message is unambiguous for those who believe that monetary policy causes inflation.<sup>9</sup>

Monetary policy, as interpreted by the Report, has sometimes been directed towards maintaining some rate of growth of the monetary stock, and at other times directed towards achieving some level of interest rates in the money markets. Moreover, doubts are expressed about the existence of evidence that a single criterion of

---

full employment in fiscal 1965 to a large full employment deficit in fiscal 1968.

<sup>8</sup> For statements consistent with this interpretation, see p. 101.

<sup>9</sup> The devaluation of the dollar can be interpreted as removing another constraint to monetary expansion. The deficit on current account and build-up of dollar balances by foreigners have inhibited monetary expansion. Hence the devaluation which will reduce, if not eliminate, the export deficit and the build-up of dollar balances removes another constraint upon monetary expansion. The Report explains the deficit in the current account as a consequence of a rise in unit labor costs between 1964 and 1969 in the United States which is twice that of our principal trading partners. See p. 151.

<sup>6</sup> The analysis of the effects of expectations of inflation contained in the Report is puzzling. Anxiety over inflation holds down business and consumer spending (p. 70) and expectations of inflation explain a building up of liquid assets (p. 110).

<sup>7</sup> It is attributed to going from a balanced budget at

monetary policy or some combination of criteria clearly stand out as being superior. To quote:

Despite the stated policy to place emphasis on the monetary growth rate in 1971, actual operations were designed to influence interest rates and conditions in short-term money markets, with the intention of thereby achieving the desired monetary growth rate. [p. 57] . . . open market operations were aimed at lowering market interest rates. [p. 57] . . . there is no single measure or objective combination of measures of monetary policy that is . . . completely satisfactory. . . . Judgment must be exercised. [p. 112]

Although the authors of the Report do not say so, it is reasonable to infer that the high growth rate of the monetary stock in recent times reflects an attempt to lower the level of interest rates. If so, then monetary policy, like wage and price controls, is being directed towards dealing with the symptoms and not the causes of inflation. High rates of growth of the monetary stock can lower interest rates in the short run, particularly short-term rates, at the expense of inflation and consequently higher future interest rates.

The Report contains an analysis of the effects of a rise in the structure or level of interest rates, presumably an unanticipated rise, upon thrift institutions and the mortgage market that is open to question. The Report argues that the yields on assets held by thrift institutions, which almost invariably are long term, generally do not rise quickly enough to enable these institutions to pay competitive rates for savings (p. 60). The authors implicitly assume that the rates thrift institutions are willing to pay for savings, setting aside regulations, is a function of past rates of return on mortgages (p. 60). Past rates of return on mortgages are irrelevant history; mortgages always yield whatever the current market is, regardless of when they were

made. An unanticipated rise in interest rates leads to unanticipated capital losses and possibly deficits in equity or capital accounts. However at the margin, what a thrift institution is willing to pay for savings is a function of what it can currently expect to earn on mortgages. Capital losses resulting from past increases in rates are irrelevant for analyzing what a thrift institution would be willing to pay for deposits when the structure of rates has risen. This can be clearly seen if one provisionally assumes free entry and asks what a new entrant would be willing to pay for deposits.<sup>10</sup>

In summary, the Economic Report of the President constitutes a defense for the imposition of direct controls and the devaluation of the dollar. The principal argument for the imposition of direct controls is that expectations of inflation must be broken. The fact of the matter is that given the rate of inflation that has occurred in recent years, expectations of inflation as revealed in our money and capital markets have probably not been strong enough, instead of too strong, as the authors contend. Moreover, expectations of inflation typically take a long time to generate, and once generated may well take a long time to eliminate. It is difficult to believe that an "emergency expedient" can play the role that the Report argues that it plays successfully.

The Report is less than candid on what explains the inflation that has occurred in the recent half-decade. Assigning the explanation to unbalanced budgets with no discussion of how these budgets were financed constitutes a highly elliptical and unsatisfactory discussion. The responsibility for the inflation that has occurred in the United States in recent years belongs

<sup>10</sup> There exist thrift institutions outside the United States whose rates paid for savings are explicitly pegged to open market rates.

in Washington and the emphasis on expectations and noncompetitive pricing as the source of inflation is misplaced. The authors never explain why monetary policy has been so expansionary although it is not hard to present an answer many find appealing by positing a tradeoff between the inflation rate and the unemployment rate. However, such a discussion requires some candor in admitting that inflation is

a cost of achieving some level of employment.

#### REFERENCES

- I. Fisher, *Theory of Interest*, New York 1965.  
W. E. Gibson, "Price-Expectations Effect on interest Rates," *J. Finance*, Mar. 1970, 25, 19-34.  
U.S. Council of Economic Advisers, *Economic Report of the President*, Washington Jan. 1972.

# The 1972 Report of the President's Council of Economic Advisers: Economics and Government

By EDMUND S. PHELPS\*

Recent revelations about the processes of government decision making in this country have revived C. P. Snow's wheezy old question of the role of science in government. It is at least timely therefore to examine the annual Report of the Council of Economic Advisers for evidence of the influence of economic science on government policy. The Council describes its annual Report as its principal means of providing the public with information and analysis of economic conditions and policies. So it seems appropriate to look there for signs of the bearing that economic knowledge and techniques have had on the important economic decisions recently taken or about to be taken by the federal government.

The conclusion of this inquiry, to arrest any suspense, is that economics seems not to have been very relevant to the recent policy decisions. The policy analysis in this year's Report is so skimpy and shallow as to project at best only the image of an assessment and defense of the government's economic policies; presumably this analysis is not substantially different from any internal analysis used by the government. The Report is more a journalistic account than what we would call a professional explanation or attempted justification of the government's policies. However disappointing a comparable review of past annual reports would be, the 1972 Report

is remarkable on three counts: (1) the assumptions about the method and effectiveness of prevailing policies that are not stated; (2) the questions about existing policy objectives that are not addressed; (3) the discussions of pending policy decisions that are not offered.

## I

The Council's Report provides a look back at the 1969-71 game plan to disinflate by means of retarding aggregate demand rather than, say, price and wage controls. There was a consensus of economists by early 1969 that the overheating of the economy had to be stopped lest the inflation rate go on rising a great deal more, if not indefinitely. The President-Elect's Task Force on Inflation recommended, as the first interim step, that aggregate demand be slowed so as to bring the unemployment rate back to some equilibrium region around 4.5 percent; one would then determine from that vantage point, in the light of whatever figure the inflation rate was beginning to steady at, the best approach to be taken to reduce the inflation rate if reduce it you must. At such a decision point it would have been reasonable to anticipate a national debate at least within the profession as to the best course of action.

What happened was that, under the cover of the expectation and acceptance of such a limited step toward reequilibration, the Administration gradually tightened

\* Professor of economics, Columbia University.

monetary and fiscal policy so severely as "gradually" to send the unemployment rate whizzing past the equilibrium zone to around 6 percent. To my knowledge the theory of how, and how well, this medicine would act to cure the patient of his inflation was never spelled out by the Council of Economic Advisers. We were never given any analysis at the time of the probable cost-effectiveness of this policy compared to variations on it (deeper and quicker, shallower and longer) and to alternative approaches such as controls or (other) structural measures on the supply side. Instead the Administration issued curious warnings that the success of the monetary and fiscal tightening in reducing inflation without side effects on unemployment would depend upon the willingness of the public to swear off price and wage increases. It was as if a physician had said that if the patient hoped to enjoy a cure, it was up to him to cooperate by responding better than others before him to the side effects of the life-threatening medication the physician had prescribed.

My point is that the Administration economists, in the Council's Report or elsewhere, did not lay out the model and the empirical estimates that would have enabled one to obtain some sense of the costs of fighting inflation in this manner—or, for a given "cost outlay," to see how rapidly inflation should be forecast to fall. This year's Report narrates how the evidence unfolding over the months of 1971 eventually tipped the scales in favor of a wage-price control program. The account here is somewhat puzzling to me. If you go by the Consumer Price Index, which is the only published index whose reciprocal purports to measure the value of money—what else is inflation about?—we find that the inflation rate from August 1969 to August 1970 was 5.6 percent. In the next twelve months, ending with the Phase I

freeze, the inflation rate fell to 4.5 percent. Even with farm-price and mortgage-rate declines removed, the deceleration of prices was proceeding about as well or better as would have been predicted from an examination of the previous inflation fight of 1957–60. The sole dark cloud that probably could not have been predicted was the failure of certain hourly compensation series to show the same speed of deceleration; indeed some series showed some acceleration in 1971. But those series are notoriously unreliable. And what if it were true that the appropriate money wage index did not decelerate at first: Are policies to be jettisoned on the first unfavorable statistical blip?

It is apparent that the record compiled by the game plan produced a public outcry and an abandonment of the plan for two reasons. The first is that the Administration's economists never confided just how bleak was the game plan's promise to begin with. The second is that evidently the Administration grossly overestimated the effectiveness of the policy. It was a case of deception and self-deception.

Suppose that the Administration's economists had candidly explained that the envisioned game plan would first have to create a disequilibrium volume of unemployment before it could generate a short-fall between actual and expected inflation rates and thus engineer a downward tendency in the latter. Suppose that the Administration's economists had predicted, on the basis of then-available evidence, that steady unemployment at the 6 percent rate would reduce the inflation rate annually by about six-tenths of a point. Then there would not have been the naive optimism that monetary policy would work some kind of magic without creating a recession and prolonged unemployment. There would either have been resignation at the slowness of progress or else an insis-

tence by the public that the game plan be supplemented or perhaps replaced by other policies from the beginning.

The advent of the multiphased wage and price control program in August of 1971 has created a new opportunity in this year's Report for a searching analysis of the method and likely effectiveness of this new policy weapon. When invited to review the Report last January, I accepted believing that it would contain some engaging ideas on this score. It was disappointing therefore to find nothing to sink any teeth into.

The Council predicts the rate of inflation over 1972 to be about 3.25 percent. This estimate, it says, "assumes that the price-wage control system, given (the conditions associated with the forecasted path of money demand), will be of the character, force, and duration needed to maintain that (forecasted) path." What does this mean? Whatever it means, 3.25 is a plausible guess. The 4.5 inflation rate of calendar 1971 was helped by a lowering of mortgage rates that, if left out, would have kept the inflation rate at nearly 5 percent. From this assumption it appears that the Council is forecasting something like a 1.75 "improvement" of the inflation rate between 1972 and 1971. The question left unanswered by the Council is what the improvement would have been, along the forecasted recovery path, had there been no wage and price controls. Six-tenths of a point at most, I would have said—especially since January 1972 may be below-normal, the bulge not having all hung out by that time. If that is right, the Council is saying that the controls will make a difference of about 1.15 for the inflation rate in 1972.

The Council's forecast of the inflation rate is essentially unexplained and, for all we know, it may be delusionary as before. The point is that there is no explicit analy-

sis of the controls policy as such. There is almost no discussion in the annual report of the determinants of the success of such control programs, no doubts raised, no issues faced. We are told that the lasting payoff of the control program is a reduction in expectations of inflation and this is to be achieved, apparently, by a demonstrable reduction of the actual rate of inflation by means of the controls. But there is little analysis of how, and to what extent, the controls developed can be counted on to accomplish that initial retardation of the rise of prices.

In their first phase, of course, the controls were comprehensive in scope. This comprehensiveness had the great merit of assuring that the general price level would in fact rise little if at all over the period of the freeze. It was hard to sympathize with the hasty critics of the freeze. To the marketeers who doubted that firms would voluntarily refrain from posting illegal price increases, one could argue that if compliance with the law is expectable in other areas, then why not in the business of price setting? To those Galbraithians who complained that the Republicans had reduced their brainchild to unworkability or absurdity, one could argue that it is only in art (and not always there) that less is more.

In Phase II the controls were limited in scope, covering less than half of the nation's value-added. At this point the likely success of the controls became more problematic. How effective (if at all) are *limited* controls in slowing the rise of the price level? If we were to postulate a model in which prices clear markets, in both equilibrium and disequilibrium, then it would seem possible to believe that the imposition of binding price ceilings on the controlled sector might induce a rise of prices in the uncontrolled sector. Yet I am not at all sure about this doctrine of spill-

over. The uncontrolled sector need not move along a rising supply curve in view of the fact that the controlled sector has to expel surplus factors of production. And even if it were true that the value of money is not really increased as a result, because there are increased waiting times to acquire goods in the controlled sector, the desired result of a reduction in the expected rate of future inflation might still follow because planned price increases have not gone through. As for ceilings on money-wage increases in the controlled sector, while they create excess demand for substitute inputs in the controlled sector, they also create excess supply for the wage-controlled inputs.

Because the Council has not bothered to raise and answer these questions the economist reading the report could be forgiven for putting little confidence in the present controls program. Indeed, one might wonder how convinced the Council is about that program.

## II

The Economic Report this year is another reflection of how often the objectives that government programs are intended to achieve are often taken for granted, evidently exempt from cost-benefit analysis. To take the conspicuous example of inflation fighting again, the Council apparently accepts as a given that inflation must be ended. It is true that we cannot spend our whole lives examining our every goal. But the uncritical acceptance by competent government economists of a policy objective is blameworthy when the costs of the programs to achieve it are clearly heavy and when the benefit from realizing the objective is not obvious and not widely agreed upon by experts.

The government economists under John Kennedy believed with fervor in the importance of raising the nation's economic growth through more austere fiscal policy

(until the unemployment problem gained precedence). If the heart had reasons, the mind never knew them. A minor intent of my parable about Oiko Nomos and his golden rule of accumulation was to spoof the certitude of those Washington economists in the optimality of faster growth. Oiko had incredible conviction, and even a theorem, but he was no more convincing than the Kennedy men.

The government's enthusiasm for growth in the early 1960's fortunately did not have serious consequences. At worst it delayed a little the support for a tax cut and delayed the eventual focus on poverty. But this is somewhat speculative. And one must esteem the sincerity of the growth advocates and envy the spirit of sacrifice in their position. There is no social goal today that calls upon everyone to give up something (unless it be space exploration). Maybe the Kennedy men were right: We want to sacrifice for the future because the people of the future are the only ones that we on this earth now can collectively sacrifice for.

The present Council under Richard Nixon has its own object of devotion—reasonable price stability. In contrast to the enthusiasm for growth, the relentless pursuit of this goal has produced the most disastrous episode of economic policy making in more than a decade. The record of rising welfare rolls, increased numbers below the poverty line, the critical curtailment of many state and local government services, and the trauma of lost jobs over the previous two years make years like 1965 or 1969 appear like glimpses of a golden age.

When the costs of the programs being applied to achieve price stability are so visibly large—and the costs of the controls program in diversions of public attention and increases of uncertainty are not negligible either—it is incumbent upon the Administration economists to make

their argument, if they have one, that the benefits of this goal exceed the costs of reaching it. The Council may say that the need for lower inflation, like the earlier need for faster growth, is self-evident. But it is not self-evident, and I suspect that this point will be as well-agreed in another ten years as the verdict today on the "growthmania" of a decade ago. It is necessary to say that the Emperor has no clothes and to demand that the Council either clothe the case for price stability in a rational cost-benefit analysis or else withdraw its allegiance to what appears to be a costly fetish.

It is true that there are some appealing and simple arguments for stability of the price level. Rising prices may confuse people and such confusion may lead to worse decisions. When prices are rising, people may, owing to their misconceived economics, feel themselves to be more "under the gun" to get wage increases; they may feel this even though the market has patently adjusted norms of wage increases. The case for steady inflation is, in contrast, more technical on most points. I will just list four arguments for inflation here and ask the reader to examine my book, *Inflation Policy and Unemployment Theory*, if he wants the details of the arguments and the way they fit together. First, the expectation of inflation levies a kind of tax which, like ordinary taxes, serves to restrain consumption and thus release resources for investment; previous writers (including myself) have emphasized the distortions of the inflation tax but not the revenue from it which lessens the need for revenue from other types of distortionary taxes. Second, an economy geared to steady inflation may be less prone to slumps and unanticipated rises of the inflation rate, because money markets are tighter and liquidity less abundant, than the same economy when it is adjusted to level prices. Third, and more

speculatively, the transition to price stability may leave permanent scars on the morale and skills of many in the labor force. Lastly, and most important, even if there were benefits to be enjoyed from adjustment to a lower rate of inflation, those benefits are spread out over the economy's whole future while the costs of getting to that allegedly happier state are bunched in the present; therefore when Doctors Burns and McCracken advocated investing for a couple of years in above-normal unemployment to bring down expectations of inflation, they should have been required to apply to the Budget Bureau with proof that the cost-benefit ratio of their investment proposal passes the Bureau's interest-rate test.

There is the distinct possibility that the Council's economists know all this and that their reply would be to say that they leave the objective of price stability unexamined because that objective is a political necessity. Washington, they might say, isn't ready for the case for inflation. Very well, then, why not let some hacks be the stooges who preside over the anti-inflation fight until Washington gets ready? The reply to that one—*we* will be better able to minimize the damage than *they* would—is possibly a snare and a delusion.

### III

I should have thought that the annual Reports of the Council would be a vehicle for debating proposals for government programs before they became policy rather than simply a forecast of what existing programs are promising. One can understand that the Administration does not want to risk public criticism of program ideas that would not have survived scrutiny in the inner councils anyway. The cost of that strategy, however, is that some objections to government policy as it is finally formulated, particularly ob-

jections from the economics profession, may not receive the attention and influence they deserve.

An example of what I have in mind is the issue of the usefulness of some form of limited price and wage controls as a permanent program. To the extent that the Phase II program of controls appears to be successful in 1972, it is foreseeable that there will be increased pressure for a permanent program of limited controls. The Council's unalterable opposition to permanent controls, indeed its distaste for controls of any kind, was set forth in the famous speech of last year by the then-chairman of the Council, Paul McCracken—a speech reprinted in the August *Monthly Economic Letter* of the First National City Bank at the very moment the Administration was initiating the new economic policy with its tight controls. This year's annual Report again dismisses the notion that controls of some kind ought to stay. The Council may well be right in its opinion but it should give its arguments, if it has any. The Council's position apparently begins and ends with an axiom that free markets are to be preferred to unfree ones whenever the inflation rate produced by the former is satisfactorily small enough. What about the counter-axiom that something or other ought to be preferred to free markets if the latter produce too much unemployment when the economy is in equilibrium at a satisfactory inflation rate?

The other example I have of the sin of omission is the failure of the annual Report even to mention the proposals being bandied about for a value-added tax. Was this not to be the year for reflection and debate on such a tax? It is true that the issue of the tax structure is not in the macro-economic policy sphere that we associate with the Council. But an annual Report that dedicated five pages to surface freight transportation could have

provided us with ten pages of background information on the tax questions that are apparently up for critical decisions in the near future.

#### IV

If the above thoughts are near the mark, they suggest that economic analysis does not occupy a strategic place in the selection of economic policy. This supports the view that has grown up from experience in a number of areas that scientific techniques and expert knowledge do not generally play an essential role in government.

In that connection an often told story is the wartime decision to carry out mass bombings of Germany. From C. P. Snow's book, *Science and Government*, and from Roy Harrod's biography, *The Prof*, we learn of the estimates by F. A. Lindemann, Churchill's personal scientific adviser, that such bombing would fairly quickly destroy half the working-class homes in the industrial towns and cities of Germany. Tizard estimated that this calculation was five times too large and Blackett, lower down in the British government, estimated it to be six times too large. Yet Lindemann prevailed. Ultimately the U.S. Survey of Strategic Bombing, a team including one J. K. Galbraith, found that Lindemann's estimates were ten times too large. Many Allied and German lives had been wasted.

The decision in 1969 of the present Administration deliberately to drive unemployment beyond the equilibrium level in order to defeat inflation bears certain similarities to that story. The effectiveness of unemployment in diminishing inflation must have been overestimated by a factor of two or three. The government could have listened to more accurate estimates based on previous evidence. Yet the game plan was quickly embraced as policy. It can only be surmised that Lindemann's calculations were accepted because his was the sort of advice that Churchill wanted

to hear. Likewise President Nixon must have been looking for an easy victory over inflation; he would believe whoever told him unemployment would be quickly efficacious toward that end. In the fight against inflation, moreover, it was never apparent why victory was important, even victory at zero cost.

No wonder that there is an absence of confidence these days in the social utility of scientific knowledge. Until ways are found of making science more immediately

effective, we have only Keynes' consolation that today's academic scribbler may shape some of the policy inclinations of politicians a generation hence.

#### REFERENCES

- R. F. Harrod, *The Prof; A Personal Memoir of Lord Cherwell*, London 1959.
- E. S. Phelps, *Inflation Policy and Unemployment Theory*, New York 1972.
- C. P. Snow, *Science and Government*, Cambridge, Mass. 1961.

# Money, Income, and Causality

By CHRISTOPHER A. SIMS\*

This study has two purposes. One is to examine the substantive question: Is there statistical evidence that money is "exogenous" in some sense in the money-income relationship? The other is to display in a simple example some time-series methodology not now in wide use. The main methodological novelty is the use of a direct test for the existence of unidirectional causality. This test is of wide importance, since most efficient estimation techniques for distributed lags are invalid unless causality is unidirectional in the sense of this paper. Also, the paper illustrates the estimation of long lag distributions without the imposition of the usual restrictions requiring the shape of the distribution to be rational or polynomial.

The main empirical finding is that the hypothesis that causality is unidirectional from money to income agrees with the postwar *U.S.* data, whereas the hypothesis that causality is unidirectional from income to money is rejected. It follows that the practice of making causal interpretations of distributed lag regressions of income on money is not invalidated (on the basis of this evidence) by the existence of "feedback" from income to money.

\* Associate professor of economics, University of Minnesota. Work for this paper was carried out during my tenure as a research fellow at the National Bureau of Economic Research, and a more extended paper on this topic may appear as a *NBER* publication. Numerous members of the *NBER* staff provided support at various stages of the research. Special thanks are due to Philip Cagan, John Hause, Milton Friedman, the Columbia Monetary Economics Workshop, and a seminar at the Cowles Foundation, whose objections and advice have sharpened the paper's argument. Josephine Su carried out the computational work. H. I. Forman drew the charts.

## I. The Causal Ordering Question for Money and Income

It has long been known that money stock and current dollar measures of economic activity are positively correlated. There is, further, evidence that money or its rate of change tends to "lead" income in some sense.<sup>1</sup> A body of macro-economic theory, the "Quantity Theory," explains these empirical observations as reflecting a causal relation running from money to income. However, it is widely recognized that no degree of positive association between money and income can by itself prove that variation in money causes variation in income. Money might equally well react passively and very reliably to fluctuations in income. Historically observed timing relations between turning points have also for some time been recognized not to be conclusive evidence for causal ordering. James Tobin and William Brainard and Tobin provide explicit examples of the possibilities for noncorrespondence between causal ordering and temporal ordering of turning points. People in close connection with the details of monetary policy know that some components of the money supply react passively to cyclical developments in the economy. Frank DeLeeuw and John Kalchbrenner, for example, argue that the monetary base (currency plus total reserves) is not properly treated as an exogenous variable in a regression equation because of the known dependence be-

<sup>1</sup> See Milton Friedman and Anna Schwartz (1963b), Friedman (1961), (1964) reprinted in chs. 10-12 of Friedman (1969).

tween certain of its components and cyclical factors.

Phillip Cagan uses an analysis of the details of money-supply determination to argue convincingly that the long-run relation between money supply and the price level cannot be due primarily to feedback from prices to money. His application of the same analytical technique to cyclical relations of money with income measures fails to yield a firm conclusion, however.

Friedman and Schwartz have argued on the basis of historical analysis that major depressions have been caused by autonomous movements in money stock.<sup>2</sup>

The issues between the monetarists and the skeptics are not easily defined on the basis of the literature cited in the preceding paragraphs. Probably few of the skeptics would deny *any* causal influence of money on income. But, on the other hand, leading exponents of the monetarist approach seem ready to admit that there is "clear evidence of the influence of business change on the quantity of money,"<sup>3</sup> at least for the mild cycles which have characterized the postwar United States.

Now if the consensus view that there is some influence of business conditions on money is correct, if this influence is of significant magnitude, and if current dollar *GNP* is a good index of business conditions,<sup>4</sup> then distributed lag regressions treating money as strictly exogenous are not causal relations. Since such regressions are now treated as causal relations by some economists, it is important to test the as-

sumption of causal priority on which they rest.

As will be shown below, there is a natural analogue in a dynamic system to Wold's "causal chain" form for a static econometric model.<sup>5</sup> This analogue turns out to be exactly a model in which causation is unidirectional according to the criterion developed by C. W. J. Granger. But Wold's form is in general not testable in a static context; any multivariate set of data with a specified list of endogenous variables can be fit by a recursive model. The dynamic analogue is, however, easily testable: If and only if causality runs one way from current and past values of some list of exogenous variables to a given endogenous variable, then in a regression of the endogenous variable on past, current, and future values of the exogenous variables, the future values of the exogenous variables should have zero coefficients.

Application of this test to a two-variable system in a monetary aggregate and current dollar *GNP* with quarterly data shows clearly that causality does not run one way from *GNP* to money. The evidence agrees quite well with a null hypothesis that causality runs entirely from money to *GNP*, without feedback.

## II. The Meaning of the Results

Before giving a rigorous explanation of the notion of causal direction and the detailed description of statistical results, it is worthwhile to consider in a nontechnical way what the results do and do not prove. That the test applied in this paper shows no feedback from  $y$  to  $x$  is a necessary condition for it to be reasonable to interpret a distributed lag regression of  $y$  on current and past  $x$  as a causal relation or to apply any of the common estimation methods involving use of lagged dependent variables or corrections for serial cor-

<sup>2</sup> See Friedman and Schwartz (1963b), p. 217-18 as reprinted in Friedman (1969).

<sup>3</sup> The quoted phrase is from Milton Friedman's introduction to Cagan, p. xxvi, and summarizes one of Cagan's main results.

<sup>4</sup> As I will argue below, it may be that the one-dimensional current dollar *GNP* index is so inadequate a measure of those aspects of business conditions which influence money supply that there is no feedback from current dollar *GNP* to money despite the existence of feedback from business conditions to money.

<sup>5</sup> See Edmond Malinvaud, p. 511 ff., for a description of causal chain models.

relation. Hence the most conservative way to state the results for money and income is that they show it to be unreasonable to interpret a least squares lag distribution for money on *GNP* as a causal relation, and that they provide no grounds for asserting that distributed lag regressions of *GNP* on money do not yield estimates of a causal relation. It is natural, and I believe appropriate, to phrase the result more positively: the data verify the null hypothesis that distributed lag regressions of *GNP* on money have a causal interpretation. However, it is possible to concoct models in which a money on *GNP* regression does not yield a causal relation and yet this paper's test would not detect feedback.

The test will fail to detect within-quarter feedback of a certain type. The "innovation" in the stochastic process  $x_t$  is that part of  $x_t$  which cannot be predicted from  $x_t$ 's own past (i.e., the residual in a regression of  $x_t$  on its own past). If  $x_t$  and  $y_t$  are connected by two causal relations—one from  $x$  to  $y$  involving a distributed lag, and the other from  $y$  to  $x$  but with only the current innovation in  $y_t$  on the right-hand side—then the test used in this paper will not detect the  $y$  to  $x$  feedback.<sup>6</sup>

Where the data show negligible serial correlation, this failing of the test becomes important. For then  $y$  and  $x$  are their own innovations and one expects that causal relations may be purely contemporaneous. In the general case, with serially correlated data, the failing is not likely to be important. It can result in false conclusions only where there is a certain sort of exact relation between the lag distributions defining the causal structure and the auto-

correlation functions of the error terms. With one important class of exceptions, there is seldom reason to suppose any relation at all between the causal structure and the properties of the error terms.

The exception arises for models in which some elements of optimal control enter. If one of the two relations in a bivariate system is chosen optimally, then the innovations in the variables become structural elements of the system. This fact is important for money and income, since it is easy to imagine that money may have been controlled to influence or to conform to income. It can be shown that in a bivariate system with optimal control of one variable, there will be in general two-way causality by the Granger criterion. The only exception is that if the information lag in the control process is just one period and if the criterion for control is minimal variance in, say,  $y$ , then causality will spuriously appear to run from  $y$  to  $x$ .<sup>7</sup> But then the only way optimal control would be likely to hide income-to-money feedback would be if income were controlled to hold down variance in money. This seems farfetched.

The fact that this paper finds no evidence of feedback from *GNP* to money is not direct evidence on the structure of money-supply determination. All that is necessary to allow interpretation of the money on *GNP* distributed lags as causal relations is the hypothesis that in this particular historical sample (1947–69), the determinants of money supply showed no *consistent* pattern of influence by *GNP*. Thus it would be enough if, for example, money supply were influenced quite differently by real and price components of *GNP* movements, so long as actual *GNP* movements were not dominated by one

<sup>6</sup> One elementary consequence is that it is possible for the test to show no feedback in either direction, despite the existence of well-defined lag distributions in both  $x$  on  $y$  and  $y$  on  $x$  regressions. This is the case where all the relation between  $y$  and  $x$  consists of contemporaneous correlation of their innovations.

<sup>7</sup> Proving this in any generality would require stretching the length and increasing the technical level of the paper. I expect to take up this point at greater length in a subsequent paper.

component or the other. Alternatively, a consistent pattern of feedback from *GNP* to money could have been swamped in this sample period by extraneous influences on money. The situation is analogous to that in a supply and demand estimation problem, where we have evidence that in a particular sample elements other than price dominated supply. Such evidence proves that in the sample the price-quantity relation traces the demand curve, but it does not in itself prove anything about the supply curve. Thus one can imagine that if heightened awareness of the importance of monetary policy makes money respond more consistently to the business cycle, single-equation estimates of the money-to-*GNP* relation will become unreliable.

Finally, we ought to consider whether the bivariate model underlying this paper could be mimicking a more complicated model with a different causal structure. The method of identifying causal direction employed here does rest on a sophisticated version of the *post hoc ergo propter hoc* principle. However, the method is not easily fooled. Simple linear structures with reversed causality like the one put forth by Tobin cannot be constructed to give apparent money-to-*GNP* causality. Complicated structures like that put forward by Brainard and Tobin in which both *GNP* and money are endogenous will except under very special assumptions yield a bivariate reduced form showing bidirectional causality. The special assumptions required to make endogenous money appear exogenous in a bivariate system must make money essentially identical to a truly exogenous variable. Thus, if money has in the sample been passively and quickly adjusted to match the animal spirits of bankers and businessmen, and if animal spirits is a truly exogenous variable affecting *GNP* with a distributed lag, then money might falsely appear to cause *GNP*.

However, if there is substantial random error in the correspondence between animal spirits and money and that error has a pattern of serial correlation different from that of animal spirits itself, then the bivariate relation between money and *GNP* will appear to show bidirectional causality.<sup>8</sup>

An assumption that future values of money or income cause current values of the other, via economic actors' having forecasts of the future better than could be obtained from current and past money and *GNP*, will affect the apparent direction of causality. However, the effect is much more likely to make a truly unidirectional structure appear bidirectional than vice versa. For example, it is easy to see that if current money supply is determined in part by extraneous knowledge of *GNP* for several future quarters, past money could spuriously appear to affect current *GNP*. However, it is difficult to imagine in such a situation why past *GNP* and all the variation in future *GNP* which can be predicted from past *GNP* should *not* affect money. Without such an artificial assumption, one cannot explain a one-sided lag distribution of *GNP* on money by a "reversed-causation-with-accurate-anticipations" model.

### III. Testing for the Direction of Causality<sup>9</sup>

In a single, static sample, the "direction of causation" connecting two related groups of variables is ordinarily not identified. That is, one can construct many different models of causal influence all of which are consistent with a given pattern

<sup>8</sup> This point is not obvious, but to prove it would, as in the case of the previous point about optimal control, overextend the paper. The technically sophisticated reader may easily verify the proposition for himself.

<sup>9</sup> It is my impression that many of the results in this section, even where they have not previously been given formal expression, are widely understood. For example, H. Akaike clearly understands that a two-sided transfer function implies the existence of feedback.

of covariances amongst the variables. If one is willing to identify causal ordering with Wold's causal chain form for a multivariate model, and if enough identifying restrictions are available in addition to those specifying the causal chain form, one can test a particular causal ordering as a set of overidentifying restrictions. The conditions allowing such a test are seldom met in practice, however.

Granger has given a definition of a testable kind of causal ordering based on the notion that absence of correlation between *past* values of one variable  $X$  and that part of another variable  $Y$  which cannot be predicted from  $Y$ 's own past implies absence of causal influence from  $X$  to  $Y$ . More precisely, the time-series  $Y$  is said to "cause"  $X$  relative to the universe  $U$  ( $U$  is a vector time-series including  $X$  and  $Y$  as components) if, and only if, predictions of  $X(t)$  based on  $U(s)$  for all  $s < t$  are better than predictions based on all components of  $U(s)$  except  $Y(s)$  for all  $s < t$ .

We will give content to Granger's definitions by assuming all time-series to be jointly covariance-stationary, by considering only linear predictors, and by taking expected squared forecast error as our criterion for predictive accuracy.

Consider the jointly covariance-stationary pair of stochastic processes  $X$  and  $Y$ . If  $X$  and  $Y$  are jointly purely linearly indeterministic (linearly regular in the terminology of Yu. S. Rozanov), then we can write

$$(1) \quad \begin{aligned} X(t) &= a^*u(t) + b^*v(t) \\ Y(t) &= c^*u(t) + d^*v(t) \end{aligned}$$

where  $u$  and  $v$  are mutually uncorrelated white noise<sup>10</sup> processes with unit variance,  $a, b, c$ , and  $d$  all vanish for  $t < 0$ , and the notation

<sup>10</sup> A "white noise" is a serially uncorrelated process.

$$g^*f(t) = \sum_{s=-\infty}^{\infty} g(s)f(t-s)$$

The expression (1) is the moving average representation of the vector process  $\begin{bmatrix} X \\ Y \end{bmatrix}$  and is unique up to multiplication by a unitary matrix.<sup>11</sup>

A useful result, not proved by Granger, is

**THEOREM 1:**  *$Y$  does not cause  $X$  in Granger's definition if, and only if,  $a$  or  $b$  can be chosen identically 0.*<sup>12</sup>

This result gives us another intuitive handle on Granger causality. If causality is from  $X$  to  $Y$  only, then of the two orthogonal white noises which make up  $X$  and  $Y$ , one is  $X$  itself "whitened" and the other is the error in predicting  $Y$  from current and past  $X$ , whitened. (A whitened variable is one which has been passed through a linear filter to make it a white noise.)

Granger has shown that if there is an autoregressive representation, given by

$$(2) \quad B^* \begin{bmatrix} X \\ Y \end{bmatrix} (t) = \begin{bmatrix} u \\ v \end{bmatrix} (t),$$

$B(t) = 0$  for  $t < 0$ ,  $u, v$  defined by (1), then the absence of causality running from  $Y$  to  $X$  is equivalent to the upper right-hand element of  $B$  being zero. That is, causality runs only from  $X$  to  $Y$  if past  $Y$  does not influence current  $X$ . From this point it is not hard to show:

**THEOREM 2:** *When  $\begin{bmatrix} X \\ Y \end{bmatrix}$  has an autore-*

<sup>11</sup> Actually, the statement that (1) is the moving average representation of  $\begin{bmatrix} X \\ Y \end{bmatrix}$  is a condition for uniqueness. There will be forms of (1) for which  $a, b, c$ , and  $d$  are all 0 for  $t < 0$  and  $u$  and  $v$  are white noises but do not yield moving average representations. These forms of (1) will not be unitary transformations of the moving average representation and can be distinguished from the true moving average representation by the fact that in a true moving average representation  $a(0)u(t) + b(0)v(t)$  is the limiting forecast error in forecasting  $X(t)$  from all past  $X$  and  $Y$ .

<sup>12</sup> Proofs of both theorems appear in the Appendix.

*gressive representation,  $Y$  can be expressed as a distributed lag function of current and past  $X$  with a residual which is not correlated with any values of  $X$ , past or future, if, and only if,  $Y$  does not cause  $X$  in Granger's sense.*

We can always estimate a regression of  $Y$  on current and past  $X$ . But only in the special case where causality runs from  $X$  to  $Y$  can we expect that no future values of  $X$  would enter the regression if we allowed them. Hence, we have a practical statistical test for unidirectional causality: Regress  $Y$  on past and future values of  $X$ , taking account by generalized least squares or prefiltering of the serial correlation in  $w(t)$ . Then if causality runs from  $X$  to  $Y$  only, future values of  $X$  in the regression should have coefficients insignificantly different from zero, as a group.

An implication of Theorem 2 is that many commonly applied distributed lag estimation techniques are valid only if causality runs one way from independent to dependent variable. The condition that the independent variable  $X$  be "strictly exogenous," central to most statistical theory on time-series regression, is exactly the Theorem 2 condition that  $X(t)$  be uncorrelated with the residual  $U(s)$  for any  $t, s$ . For example, quasi differencing to eliminate serial correlation in residuals will produce inconsistent estimates without the one-way causality condition; and the "Koyck transformation" which is invoked to allow interpretation of regressions with autoregressive terms as estimates of infinite lag distributions depends on one-way causality. Hence in principle a large proportion of econometric studies involving distributed lags should include a preliminary test for direction of causality.

#### *Remarks on Distributed Lag Methodology*

Especially in a study of this kind, where we wish to make fairly precise use of  $F$ -

tests on groups of coefficients, it is important that the assumption of serially uncorrelated residuals be approximately accurate. Therefore all variables used in regressions were measured as natural logs and prefiltered using the filter  $1 - 1.5L + .5625L^2$ ; i.e., each logged variable  $x(t)$  was replaced by  $x(t) - 1.5x(t-1) + .5625x(t-2)$ . This filter approximately flattens the spectral density of most economic time-series, and the hope was that regression residuals would be very nearly white noise with this prefiltering.

Two problems are raised by this prefiltering. First, if the filter has failed to produce white noise residuals, it is quite unlikely to fail by leaving substantial positive first-order serial correlation. Durbin-Watson statistics are therefore of little use in testing for lack of serial correlation, and tests based on the spectral density of the residuals were used instead. Second, as I pointed out in an earlier paper (1970), prefiltering may produce a perverse effect on approximation error when lag distributions are subject to prior "smoothness" restrictions. Therefore, no Koyck, Almon, or rational lag restrictions were imposed a priori, and the length of the estimated lag distributions was kept generous.

In applying the  $F$ -tests for causal direction suggested in the previous section, one should bear in mind that the absolute size of the coefficients is important regardless of the  $F$  value. It is a truism too often ignored that coefficients which are "large" from the economic point of view should not be casually set to zero no matter how statistically "insignificant" they are. Thus, the fact that future values of the independent variable have coefficients insignificantly different from zero only shows that unidirectional causality is possible. If the estimated coefficients on future values are as large or larger than those on past values, bidirectional causality may be very important in practice, despite in-

significant  $F$ 's. Moreover, small coefficients on future values of the independent variable may sometimes be safely ignored even when they are statistically significant. This is especially true in the light of my observation (1971) that nonzero coefficients on future values may be generated in discrete-time data from a "one-sided" continuous-time distributed lag.<sup>13</sup>

All the data used in the regressions presented in this paper were seasonally adjusted at the source. This creates potential problems of a sort which has not been widely recognized heretofore. Most seasonal adjustment procedures in common use allow for a seasonal pattern which shifts slowly over time, and the rate at which the seasonal pattern is taken to shift varies from one series to another. It can be shown<sup>14</sup> that in distributed lag regressions relating two variables which have been deseasonalized by procedures with different assumed rates of shift in the seasonal pattern, spurious "seasonal" variation is likely to appear in the estimated lag distribution. The lag distributions estimated in this paper are long enough and free enough in form that bias from this source should be obvious wherever it is important (and it is important in one regression). However, it would be better to start from undeseasonalized data, being sure that both variables in the relation are deseasonalized in the same way. A check along these lines, using frequency-domain procedures, was carried out for this paper and is mentioned in the discussion of results below.

<sup>13</sup> The definition of causality given in the previous section generalized easily to continuous time. One simply reinterprets (1) as a continuous-time relation, and " $Y$  does not cause  $X$ " still corresponds to " $b$  identically zero."

<sup>14</sup> I showed this in an earlier mimeographed version of this paper. A separate short paper on this topic is in preparation.

#### IV. Time Domain Regression Results

The data used cover the period 1947-69, quarterly. Money was measured both as monetary base ( $MB$ )—currency plus reserves adjusted for changes in reserve requirements—and as  $M1$ —currency plus demand deposits. Figures for  $MB$  were taken from the series prepared by the Federal Reserve Bank of St. Louis and supplied to the National Bureau of Economic Research data bank. Results were similar for  $M1$  and  $MB$ , so we sometimes use  $M$  or money to refer to both  $M1$  and  $MB$  in what follows.

Regressions of the  $\log$  of  $GNP$  (in current dollars) on future and lagged  $\log M$  were significant, as were the reversed regressions of  $\log M$  on future and lagged  $\log GNP$ . (See Table 1.) Table 2 reports tests for homogeneity between the pre-1958 and post-1958 sections of the sample. No significant differences between the subsamples appeared in the regressions. Future values of  $GNP$  were highly significant in explaining the  $M$  dependent variable, but future values of  $M$  were not significant in explaining the  $GNP$  dependent variable. (See Table 3.) The largest individual coefficients in each  $GNP$  on  $M$  regression occur on past lags,

TABLE 1—SUMMARY OF OLS REGRESSIONS\*

	$F$ for Independent Variables	$\bar{R}^2$	Standard Error of Estimate	Degrees of Freedom
$GNP=f(M1, 8 \text{ past lags})$	1.89*	0.7927	0.01018	64
$GNP=f(M1, 4 \text{ future, 8past lags})$	1.37	0.7840	0.01040	60
$GNP=f(MB, 8 \text{ past lags})$	2.24**	0.8004	0.00999	64
$GNP=f(MB, 4 \text{ future, 8past lags})$	1.61	0.7924	0.01019	60
$M1=f(GNP, 4 \text{ future, 8past lags})$	11.25**	0.8385	0.00403	60
$MB=f(GNP, 4 \text{ future, 8past lags})$	5.89**	0.8735	0.00410	60

\* Significant at 0.10 level.

\*\* Significant at 0.05 level.

\* All regressions were fit to the period 1949III-1968IV.  $M1$  is currency plus demand deposits.  $MB$  is monetary base as prepared by the Federal Reserve Bank of St. Louis. The  $F$ -tests shown are for the null hypothesis that all right-hand side variables except trend and seasonal dummies had zero coefficients. See also notes to Table 4.

TABLE 2—*F*'S FOR COMPARISONS OF SUBPERIODS  
1948III–1957III vs. 1957IV–1968IV<sup>a</sup>

Regression Equation	<i>F</i>	Degrees of Freedom
$GNP = f(M1, 8 \text{ past lags})$	1.44	(14, 50)
$GNP = f(MB, 8 \text{ past lags})$	0.64	(14, 50)
$M1 = g(GNP, 4 \text{ future, 8 past lags})$	0.88	(18, 46)
$MB = f(GNP, 4 \text{ future, 8 past lags})$	1.01	(18, 46)

<sup>a</sup> Tests are for the null hypothesis that all coefficients (including trend and seasonals) remained the same in both subsamples.

and the estimated shapes for those regressions appear broadly reasonable on the assumption that coefficients on future lags are small and coefficients on past lags are nonzero and fairly smooth. (See Table 4 and Figures 1 and 2.)

These results allow firm rejection of the hypothesis that money is purely passive, responding to *GNP* without influencing it. They are consistent with the hypothesis that *GNP* is purely passive, responding to

TABLE 3—*F*-TESTS ON FOUR FUTURE  
QUARTERS' COEFFICIENTS<sup>a</sup>

Regression Equation	<i>F</i>
$GNP \text{ on } M1$	0.36
$GNP \text{ on } MB$	0.39
$M1 \text{ on } GNP$	4.29**
$MB \text{ on } GNP$	5.89**

\*\* Significant at 0.05 level

<sup>a</sup> All tests apply to regressions run over the full sample and are assumed distributed as  $F(4, 60)$ .

*M* according to a stable distributed lag but not influencing *M*.

But let us note a few statistical *caveats*. Though the estimated distribution looks like what we expect from a one-sided true distribution, the standard errors on the future coefficients are relatively high. These results are just what a unidirectional causality believer would expect, but they are not such as to necessarily force a believer in bidirectional causality to change his mind. Also, seasonality problems are

TABLE 4—LAG DISTRIBUTIONS FROM TIME-DOMAIN REGRESSIONS<sup>a</sup>

Coefficient on lag of:	<i>GNP</i> on <i>MB</i> past only	<i>GNP</i> on <i>MB</i> with future	<i>MB</i> on <i>GNP</i>	<i>GNP</i> on <i>M1</i> past only	<i>GNP</i> on <i>M1</i> with future	<i>M1</i> on <i>GNP</i>
−4		−0.65	.162		−.300	.050
−3		.290	−.013		.120	.117
−2		−.088	.105		.126	.069
−1		−.110	.179		.105	.125
0	.603	.532	.171	.570	.484	.181
1	.593	.507	.015	.370	.412	.089
2	.509	.515	.052	−.034	−.017	.116
3	−.029	.080	.264	.543	.582	.107
4	−.011	.023	.107	−.242	−.363	.027
5	−.865	−.822	−.009	−.178	−.147	.027
6	−.037	−.053	.016	−.180	−.136	.025
7	−.296	−.282	.147	−.157	−.139	.123
8	.072	.039	.130	−.326	−.405	.112
Standard errors of coefficients:						
Largest s.e.	.313	.338	.052	.293	.318	.051
Smallest s.e.	.272	.276	.045	.274	.294	.044
Sum of coefficients	.540	—	—	.365	—	—
Standard error of sum	.442	—	—	.523	—	—

<sup>a</sup> Regressions were on *logs* of variables, prefiltered as explained in the text. Each regression included, in addition to the leading and lagging values of the independent variable for which coefficients are shown, a constant term, a linear trend term, and three seasonal dummies. Trends were in all cases significant. Seasonal dummies were insignificant. (The data were seasonally adjusted.)

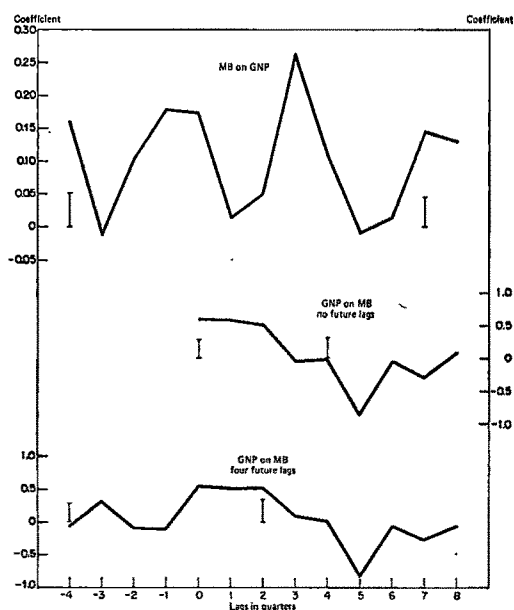


FIGURE 1. LAG DISTRIBUTIONS FOR *MB* AND *GNP*  
 Note: Smallest and largest standard errors are displayed as vertical lines above or below corresponding coefficients.

clearly present in the *MB* on *GNP* regression. Seasonality effects appear to be less of a problem with *M1* than with *MB*.

DeLeeuw and Kalchbrenner have argued, in attacking the "reduced form" money vs. *GNP* regressions put out by the St. Louis Fed, that the monetary base is not truly exogenous. We have discussed above the substance of that argument. Suffice it to say here that they claim that one could make the monetary base more "exogenous" by extracting from it borrowed reserves and (possibly) cash in hands of the public. Attempts to use these adjusted *MB* series (one of them is actually unborrowed reserves) failed, in the sense that relations were less significant statistically and *GNP* on adjusted *MB* regressions did not show one-sided lag distributions.

The same regression equations used for *GNP* and *M* were estimated also with *GNP* replaced by the *GNP* deflator

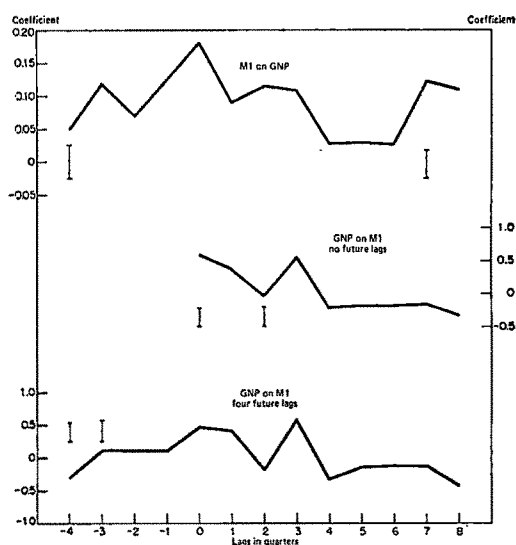


FIGURE 2. LAG DISTRIBUTIONS FOR *M1* AND *GNP*  
 Note: Smallest and largest standard errors are displayed as vertical lines above or below corresponding coefficients.

(*PGNP*) and then by real *GNP* (*RGNP*) with *MB* the money variable. Quantity theory even in its modern guise does not claim to have firm implications about the way income changes divide into real and price components, but it seemed useful to examine the possibility that monetary variables would predict the components separately as well as their product. Standard errors of the (logarithmic) equations regressing *RGNP* on *MB* were slightly larger than corresponding standard errors for current dollar *GNP*. Values of coefficients and *F*-statistics were much the same with *RGNP* as dependent variable as with *GNP* the dependent variable. Future lags were again highly significant for *MB* on *GNP* regressions and highly insignificant for the reversed relation. However, with *RGNP*, current plus eight past lagged values of *MB* were not as a group significantly different from zero at the .10 level. With *PGNP*, standard errors of estimate were small, but almost every *F*-test failed to attain significance, in-

cluding the test on future lags in the *MB* on *PGNP* relation.

## V. Tests for Serial Correlation in Residuals

Durbin-Watson statistics for all reported regressions are close to two. This is to be expected because of the pre-filtering. The test on the cumulated periodogram of the residuals, described by James Durbin, yields results in the indeterminate range for each regression.<sup>15</sup> The test on the cumulated periodogram is in principle capable of detecting departures from serial independence even when there is no first-order serial correlation, and in this sense is a stronger test than the Durbin-Watson for the case at hand.

The central difficulty here, though, is that a total of 17 of the available 78 degrees of freedom have been used up in the regression, so that the easily-computed bounds tests leave a wide range of indeterminacy. An alternative to the bounds tests is to use the likelihood ratio test for the null hypothesis that the periodogram of the residuals has constant expectation across a number of intervals. This test is described in E. J. Hannan (1960), p. 98.<sup>16</sup> In application to regression residuals this test is justified only when the number of observations is much larger than the number of independent variables, which is clearly not the case here. The statistics reported in Table 5 would be distributed as chi-square with 7 degrees of freedom if asymptotic results applied, but the true

TABLE 5—LIKELIHOOD-RATIO TESTS FOR WHITE NOISE RESIDUALS<sup>a</sup>

<i>GNP</i> on <i>MB</i>	<i>GNP</i> on <i>M1</i>	<i>MB</i> on <i>GNP</i>	<i>M1</i> on <i>GNP</i>
13.02	19.01	11.04	12.64

Note: .05 significance level for chi-squared with 7 degrees of freedom. 14.1

<sup>a</sup> The statistics shown are each distributed asymptotically as chi-square with 7 degrees of freedom on the null hypothesis of white noise residuals. As noted in the text, the asymptotic distribution is probably not a good approximation to the true distribution here. For the *GNP* on *M* equations, residuals were taken from the form with no future lags. For the *M* on *GNP* equations, residuals were taken from the form including future lags.

significance levels of the test will be higher than the nominal ones. Even at nominal significance levels, though, only the residuals from the regression of *GNP* on *M1* are significantly "nonwhite" at a 5 percent level.

The conclusion from this list of approximate or inconclusive tests can only be that there is room for doubt about the accuracy of the *F*-tests on regression coefficients.

As a check on the least squares results, these same regressions were estimated also using a frequency-domain procedure, Hannan's (1963) "inefficient" procedure.<sup>17</sup> This procedure has some disadvantages relative to least squares, but it has the two advantages that 1) it makes it computationally simple to estimate the variance-covariance structure of the residuals and use the estimate in constructing tests on the estimated regression coefficients and 2) it makes it easy to deseasonalize raw data directly. Not all the tests for significance of groups of coefficients came out

<sup>15</sup> The test carried out was actually based on cumulation of the periodogram over 128 equally spaced points, instead of over the 39 harmonic frequencies as would be appropriate to get Durbin's test. This difference is, however, demonstrably asymptotically negligible (as sample size increases Durbin's test converges in distribution to any test based on more points than half the sample size) and seems unlikely to have been very important even at this particular sample size.

<sup>16</sup> Hannan's description includes Bartlett's small-sample correction to the likelihood ratio test. The results reported in Table 5 do not include the Bartlett correction, since it was small.

<sup>17</sup> The theory of these estimates has been extended in Hannan (1967) and Wahba. It is worthwhile noting that Wahba's proof that the Hannan inefficient estimates are "approximately" least squares estimates is not a proof that the Hannan inefficient estimates have the same asymptotic distribution as least squares, and their asymptotic distributions are in fact different.

the same way at the same significance levels in the frequency-domain estimates, but the general agreement with the least squares results was so close that there is no point in reproducing the frequency-domain results here.<sup>18</sup> Raw data for the monetary base was not readily available, but frequency-domain estimates using raw data on *M1* and *GNP*, symmetrically deseasonalized, gave results very similar to those obtained with least squares on published deseasonalized data.

## VI. The Form of the Lag Distribution

The lag distribution estimated here to relate *GNP* to *M* has only a loosely determined form because of the lack of prior restrictions on its shape. Still, it is worthwhile noting that it agrees in general shape with many previous estimates, and that it can be given an economic explanation. The distribution is positive at first, then becomes mostly negative beyond the fourth lag. The initial positive coefficients sum to a number greater than one, though the sum of all the coefficients is less than one. (Note, though, that the standard error on the sum of coefficients is very large. See Table 4.) The pattern of a short-run elasticity exceeding unity and a long-run elasticity below unity agrees with the theoretical speculations of Friedman (1969), pp. 138–39, concerning the effects of a demand for money dependent on permanent rather than on current income. However, note that the contemporaneous quarter response is less than unitary, and that negative response does not set in for several quarters. To explain this, one must either introduce an averaging procedure into the other side of the equation, making “permanent money” depend on permanent income, or one must introduce the possibility of transactional fric-

tions which keep the economy off its demand curve for money in the short run. At least the latter of these elements is not novel. Alan Walters pointed out that over short enough time intervals people are likely to be off their demand curves. It seems only natural that, since individuals' money balances always fluctuate over short periods due to random timing of transactions, it should take time for changes in money balances to affect individuals' spending behavior.

## VII. Conclusion

The main conclusions of the paper were summarized in the introduction. I repeat them more briefly here: In time-series regression it is possible to test the assumption that the right-hand side variable is exogenous; thus the choice of “direction of regression” need not be made entirely on a priori grounds. Application of this test to aggregate quarterly data on *U.S. GNP* and money stock variables shows that one clearly should not estimate a demand for money relation from these data, treating *GNP* as exogenous with money on the left-hand side; no evidence appears to contradict the common assumption that money can be treated as exogenous in a regression of *GNP* on current and past money.

## APPENDIX

**THEOREM 1:** *Y does not cause X in Granger's definition if, and only if, in the moving average representation*

$$(A) \quad \begin{bmatrix} X(t) \\ Y(t) \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} * \begin{bmatrix} u \\ v \end{bmatrix} (t),$$

*a or b can be chosen to be identically zero.*

## PROOF:

Following Rozanov we introduce the notation  $H_s(t)$  to stand for the completion under the quadratic mean norm of the linear space of random variables spanned by  $z(s)$  for  $s \leq t$ . Suppose  $b$  is zero. Clearly  $X(t)$  then

<sup>18</sup> The frequency-domain results were presented and discussed in an earlier mimeographed version of this paper.

lies in  $H_u(t)$ . By the definition of a moving average (*m.a.*) representation,  $H_{X,Y}(t)$  is identical to  $H_{u,v}(t)$ . But it follows from Rozanov's "Remarks" on pages 62-63 that if  $H_u(t)$  and  $H_X(t)$  are not identical, then with  $b$  zero the identity of  $H_{X,Y}(t)$  and  $H_{u,v}(t)$  fails. Therefore,  $H_u(t)$  and  $H_X(t)$  are identical. But then the projection of  $X(t)$  on  $H_{X,Y}(t-1)$  is in  $H_X(t-1)$ , which is to say that given past  $X$ , past  $Y$  does not help in predicting current  $X$ . One side of the double implication is proved.

In Granger's definition,  $Y$  not causing  $X$  is the same thing as the projection of  $X(t+1)$  on  $H_{X,Y}(t)$  lying in  $H_X(t)$ . Assuming this condition holds, define  $u(t)$  as the difference between  $X(t)$  and the projection of  $X(t)$  on  $H_X(t-1)$ . Define  $w(t)$  as the difference between  $Y(t)$  and the projection of  $Y(t)$  on  $H_{X,Y}(t-1)$ . Finally, define  $v(t)$  as that part of  $w(t)$  orthogonal to  $u(t)$  (i.e., the residual in a regression of  $w(t)$  on  $u(t)$ ). By definition,  $u(t)$  and  $w(t)$  and therefore,  $v(t)$  are uncorrelated with past values of each other. Also,  $u(t)$  and  $v(t)$  are contemporaneously uncorrelated and  $H_{u,v}(t)$  is identical to  $H_{X,Y}(t)$ . Expressing  $X(t)$  and  $Y(t)$  in terms of the coordinates  $u(s)$ ,  $s \leq t$ , will give us a moving average representation of the form (A).

**THEOREM 2:** *When  $\begin{bmatrix} X \\ Y \end{bmatrix}$  has an autoregressive representation,  $Y$  can be expressed as a distributed lag function of current and past  $X$  with a residual which is not correlated with any  $X(s)$ , past or future if, and only if,  $Y$  does not cause  $X$  in Granger's sense.*

**PROOF:**

Suppose  $Y$  can be expressed as a distributed lag on  $X$  with a residual  $w(t)$  independent of  $X(s)$  for all  $s$ . Let  $u(t)$  be the fundamental white noise process in the moving average representation of  $X(t)$  alone and  $v(t)$  be the fundamental white noise process in the *m.a.* representation of  $w(t)$  alone. Write the assumed distributed lag relation

$$(B) \quad Y(t) = \mu^* X(t) + w(t)$$

Then clearly

$$(C) \quad Y(t) = \mu^* a^* u(t) + d^* v(t),$$

where  $a^* u$  and  $d^* v$  are the *m.a.* representations of  $X$  and  $w$ , respectively. The equation (C) together with the *m.a.* representation of  $X$  are clearly in the form (A) with  $b \equiv 0$ . Now we need only verify that  $u$  and  $v$  are jointly fundamental for  $X$  and  $Y$ , and for this we need only show that  $H_{X,Y}(t)$  includes  $H_{u,v}(t)$ .  $H_u(t)$  is in  $H_X(t)$  by definition.  $H_v(t)$  is in  $H_w(t)$  which is in turn (by inspection of (B)) in  $H_{X,Y}(t)$ . One side of the double implication is proved. Suppose that we have the autoregressive representation

$$(D) \quad \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix} * \begin{bmatrix} X \\ Y \end{bmatrix} (t) = \begin{bmatrix} u \\ v \end{bmatrix} (t)$$

and that the *m.a.* representation has the form (A) with  $b \equiv 0$ . Let  $G$  be the matrix on the right-hand side of (A) and  $H$  be the matrix on the left-hand side of (D). Then almost everywhere  $\tilde{G}^{-1} = \tilde{H}$ . (The tilde denotes a Fourier transformation.) Since  $\tilde{G}$  can be written in triangular form,  $\tilde{H}$  (and thus  $H$ ) can be written triangular also. But then we can substitute the first equation of (D) into the second equation of (A) to obtain

$$(E) \quad Y(t) = c^* \alpha^* X(t) + d^* v(t)x$$

Equation (E) has the desired properties, since  $X$  can be expressed entirely in terms of  $u$  and  $v$  is uncorrelated with  $u$ .

## REFERENCES

- H. Akaike, "Some Problems in the Application of the Cross-Spectral Method," in B. Harris, ed., *Advanced Seminar on Spectral Analysis of Time Series*, New York 1967.
- W. Brainard, and J. Tobin, "Pitfalls of Financial Model Building," *Amer. Econ. Rev. Proc.*, May 1968, 58, 99-122.
- P. Cagan, *Determinants and Effects of Changes in the Stock of Money*, Nat. Bur. Econ. Res. Stud. *Business Cycles*, No. 13, New York 1965.
- F. DeLeeuw, and J. Kalchbrenner, "Monetary and Fiscal Actions: A Test of Their Relative Stability—Comment," *Fed. Reserve Bank St. Louis Rev.*, Apr. 1969, 51, 6-11.
- J. Durbin, "Tests for Serial Correlation in Regression Analysis Based on the Periodo-

- gram of Least Squares Residuals," *Biometrika*, Mar. 1969, 56, 1-16.
- M. Friedman, "The Monetary Studies of the National Bureau," Nat. Bur. Econ. Res. *Annual Report* 1964; reprinted in Friedman (1969).
- , "The Lag in the Effect of Monetary Policy," *J. Polit. Econ.*, Oct. 1961, 69, 447-66; reprinted in Friedman (1969).
- , *The Optimum Quantity of Money and Other Essays*, Chicago 1969.
- and A. Schwartz, (1963a) *Monetary History of the United States 1867-1960*, Nat. Bur. Econ. Res. *Stud. Business Cycles*, No. 12, Princeton 1963.
- and ———, (1963b) "Money and Business Cycles," *Rev. Econ. Statist.*, Feb. 1963, supp., 45, 32-64; reprinted in Friedman (1969).
- C. W. J. Granger, "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods," *Econometrica*, July 1969, 37, 424-38.
- E. J. Hannan, *Time Series Analysis*, London 1969.
- , "Regression for Time Series," in M. Rosenblatt, ed., *Time Series Analysis*, New York 1963.
- , "Estimating a Lagged Regression Relation," *Biometrika*, 1967, 54, 409-18.
- E. Malinvaud, *Statistical Methods of Econometrics*, Chicago 1965.
- Y. S. Razanov, *Stationary Random Processes*, San Francisco 1967.
- C. A. Sims, "The Role of Approximate Prior Restrictions in Distributed Lag Estimation," *J. Amer. Statist. Ass.*, Mar. 1972, 67, 169-75.
- , "The Role of Approximate Prior Restrictions in Distributed Lag Estimation," mimeo 1970.
- , "Discrete Approximation to Continuous-Time Distributed Lags in Econometrics," *Econometrica*, May 1971, 39, 545-63.
- J. Tobin, "Money and Income: Post Hoc Ergo Propter Hoc?" *Quart. J. Econ.*, May 1970, 84, 301-17.
- G. Wahba, "Estimation of the Coefficients in a Multi-Dimensional Distributed Lag Model," *Econometrica*, July 1969, 37, 398-407.
- A. Walters, "Professor Friedman on the Demand for Money," *J. Polit. Econ.*, Oct. 1965, 73, 545-55.

# Neoclassical Investment Models and French Private Manufacturing Investment

By RICHARD SCHRAMM\*

A major development in economics during the last ten years was the joining of the neoclassical theory of capital with the econometric study of investment behavior.<sup>1</sup> While there is controversy over the interpretation of some of the empirical results of neoclassical investment demand equations,<sup>2</sup> this joining of theory and empirical study has led to increased emphasis on the derivation and specification of testable models and more frequent

inclusion of output price and factor cost variables in investment studies.<sup>3</sup>

Available data on French investment since 1950 provides a useful sample to test neoclassical investment models in a different economic setting. The many differences between the *U.S.* and French economy in both structure and economic experience since 1950 allow tests of both the theoretical flexibility and the empirical validity of investment models which have proved generally successful in their application to *U.S.* experience.

This study employs the neoclassical theory of capital to derive a model of the demand for investment goods by a firm in France. The model relates the net present value of the firm to output and factor market variables. It incorporates the major characteristics of the French business tax system and changes in this system since 1950, different prices subject to credit and anti-inflation policies, and product demand factors. Using these relationships, I derive models of industry investment and test these models using regression analysis of annual time-series data, 1950-65, for the private manufacturing sector of France and eight industry subgroups of this sector.

## I. The French Firm's Environment for Investment Decision Making

As a basis for investment analysis, this section derives the relationship be-

\* Cornell University. This study was conducted under a Fulbright Fellowship in France and benefited from discussions with numerous French academic and government officials. In particular I would like to thank J. Mairesse of the Institut National de la Statistique et des Études Économiques for his valuable assistance. The initial research received support from the Faculty Research Fund of the Graduate School of Business, Columbia University, with computer time provided by IBM France and Columbia University. The final phases of the study were supported by funds from the Graduate School of Business and Public Administration, Cornell University. Roger Sherman, Maurice Wilkinson, Jerome Hass, Seymour Smidt and Robert Tollison provided very useful comments on earlier drafts of the paper and Ricardo Sanchez-Aguilar furnished much appreciated assistance with the empirical analysis.

<sup>1</sup> Contributions during this period include the theoretical developments of Trygve Haavelmo, their refinement and application to the econometric study of investment behavior by Dale Jorgenson, the adaptation of factor cost variables to include income tax parameters by Jorgenson, Robert Hall and Jorgenson, and Robert Coen (1968), and the joining of neoclassical theory and accelerator investment models by Robert Eisner and Robert Strotz. Econometric studies building on these developments include those by Jorgenson and others (e.g. 1963, 1967a, 1967b, 1968, 1969), Bert G. Hickman, and Schramm.

<sup>2</sup> See, for example Hall and Jorgenson (1967, 1969), Eisner and M. I. Nadiiri (1968, 1970), Charles Bischoff (1969), Jorgenson and J. Stephenson (1969), Eisner (1969), Coen (1969), and J. C. R. Rowley.

<sup>3</sup> See, for example, Bischoff, Coen (1968), Hickman, and Schramm, and the references in fn. 2.

tween the level of capital stock held by the firm and the net present value of future returns to the firm from this capital investment. These future returns depend upon output and factor market conditions, including a variety of tax effects, and production and capacity considerations. These relationships are used in Section II to derive and test models of investment at the industry level.

Assume the firm produces one homogeneous product,  $Q$ , by combining capital services from its stock of capital,  $K$ , and labor services,  $L$ , according to the production function

$$(1) \quad Q_t = Q(K_t, L_t)$$

This production process will be assumed to represent output  $Q_t$  as occurring over the period  $t$  from the combination of the capital stock  $K_t$  and labor flow  $L_t$  existing at the beginning of the period  $t$ . For later considerations of discounting revenue streams, all transactions for output, labor, and investment goods will be assumed to occur at the end of the period. Net additions to capital stock,  $K_{t+1} - K_t$ , occurring during the period  $t$  will change the level of initial capital stock for the next period but not affect current production. Gross investment over  $t$  will be represented as

$$(2) \quad I_t = K_{t+1} - K_t + \delta K_t$$

with the capital replacement rate  $\delta$  treated as constant over the period under consideration.

Operating subject to the above production relationships, the producer faces the corresponding markets for output, labor, capital equipment, as well as a market for capital funds. It is assumed the producer knows the actual prices prevailing in these markets up to the beginning of the period  $t$  and that he forms constant price expectations in all these markets over his planning horizon. These expected prices will be rep-

resented as  $p$ ,  $w$ ,  $q$ , and  $r$  for the output, labor, capital equipment, and capital funds markets, respectively.

The fiscal influences of the French government incorporated in the model are the change in the tax structure in 1954-55 (including the resulting changes in capital asset valuation), the changes in the rates of different taxes affecting business operations over the entire period, and the change in depreciation procedures for tax purposes in 1959-60. These changes can all be incorporated in the representation of total direct and indirect taxes paid by the firm each period.

The three major taxes paid by French firms are the tax on value-added, taxes associated with total wage payments, and taxes on corporate income (similar to the U.S. corporate income tax). The tax on value-added,  $T_v$ , is<sup>4</sup>

$$(3) \quad T_v = t_v p Q - v t_i q I$$

— other value-added tax payments

where  $t_v$  is the value-added tax rate applied to total sales,  $t_i$  is the rate of value-added taxes paid on investment goods purchased, and  $v$  is the proportion of the taxes paid in purchasing investment goods which are deductible from  $T_v$ . Deductions of other value-added tax payments, such as for intermediate goods, have not been included specifically since their effect on investment behavior is probably very small.

This expression for total value-added taxes incorporates not only the value-added tax rate paid on sales and on investment goods but also, through the introduction of the parameter  $v$ , the structural

<sup>4</sup> In France value-added tax payments to the government represent the difference between taxes owed based on the value of domestic sales and the value-added taxes paid by the firm on certain goods used in production. For a general description of the French system, see Clara Sullivan.

change introduced in the tax system in 1954-55. Prior to this period, taxes paid on investment goods were not deductible from taxes on production (i.e.,  $v=0$ ); afterwards taxes on most investment expenditures could be deducted (i.e.,  $v \approx 1$ ). In addition this change required that capital assets be valued on an after-tax basis for depreciation purposes.

Taxes associated with wage payments,  $T_s$ , include a tax levied directly on total salaries paid and social security payments. These tax payments each period can be combined and approximated as

$$(4) \quad T_s = t_s wL$$

where  $t_s$  is the "effective" wage tax rate.

Direct taxes  $T_b$  paid by private firms consist of taxes on corporate income. Most large firms in the private sector are subject to direct taxation of this form which can be approximated as

$$(5) \quad T_b = t_b(pQ - wL - D - T_v - T_s - \text{other deductions})$$

where  $t_b$  is the tax rate on corporate income and  $D$  is the depreciation allowed for tax purposes. Other deductions include purchases of materials and intermediate goods.

Depreciation allowances affect total tax payments through their deductibility from direct tax payments as indicated by equation (5) above. Up to 1960, firms could employ only straight line methods of calculating capital asset depreciation for tax purposes. Beginning with investments in 1960, however, a form of declining balance depreciation was allowed. In each period investment  $I_t$  generates tax deductions over the life of the investment based on the after-value-added-tax value of the investment,  $(1-vt_i)qI_t$ , and the depreciation allowance scheme being followed. Assuming all firms made this changeover in 1960, depreciation  $D_t$  can be represented as:

$$(6a) \quad D_t = \sum_{j=t-N}^{j=t} \frac{1}{N} q_i(1-vt_i)_i I_j \quad \text{for } t < 1960$$

$$(6b) \quad D_t = \sum_{j=t-N}^{j=1959} \frac{1}{N} q_i(1-vt_i)_i I_j + \sum_{j=1960}^{j=t} \frac{a}{N} \left(1 - \frac{a}{N}\right)^{t-j} q_i(1-vt_i)_i I_j \quad \text{for } t \geq 1960$$

where  $N$  is the average life of depreciable assets and  $a$  is the average declining balance depreciation rate.

Under these methods of depreciation allowances, the present value of the depreciation deductions arising from one monetary unit of investment is

$$(7a) \quad Z_t = \frac{1}{N} \sum_{j=1}^{j=N} (1+r)^{-j} \quad \text{for } t < 1960$$

$$(7b) \quad Z_t = \sum_{j=1}^{j=N} \frac{a}{N} \left(1 - \frac{a}{N}\right)^{j-1} (1+r)^{-j} \quad \text{for } t \geq 1960$$

Treating price and tax parameters as long-run expected constant levels, the present value in period  $t$  of the deductions allowed under the two depreciation systems in France is

$$(8) \quad D'_t = (1-vt_i)qI_t Z_t$$

Government monetary policies are assumed to influence the investment behavior of firms through their effect on the firm's expected cost of capital,  $r$ . Government monetary activities related to investment behavior in France are quite complex, ranging from the direct or indirect supply of investment funds to firms to government market behavior affecting rates in private financial markets. As varied as these activities are, it will be assumed here that these actions have the ultimate effect of influencing the cost of

investment funds to firms reflected in private bond market rates.

In developing this theory of firm behavior, the device which serves to pull together production relationships, market conditions, and government monetary and fiscal policy variables is the assumption that the firm incorporates all these considerations in its assessment of the influence of production and investment decisions on the present value of future cash flows.

The expected after-tax cash flows in period  $t$  are

$$(9) \quad R_t = pQ_t - wL_t - qI_t \\ - (T_v + T_b + T_s)$$

The only tax effects from period  $t$  decisions that occur in other periods result from depreciation expense associated with  $I_t$ . To bring all tax effects from  $Q_t$ ,  $L_t$ , and  $I_t$  into period  $t$ , a modified cash flow expression,  $R'_t$ , can be constructed which excludes depreciation effects from previous investments but includes the present value of future depreciation deductions,  $D'_t$ , arising from  $I_t$ .

Substituting tax equations (3), (4), and (5) and equation (8) into equation (9) and simplifying we have

$$(10) \quad R'_t = p'Q_t - w'L_t - q'I_t$$

where

$$p' = (1 - t_b)(1 - t_v)p$$

$$w' = (1 + t_s)(1 - t_b)w$$

$$q' = [1 - t_v(1 - t_b) - t_b(1 - v t_i)Z]q$$

Under these conditions the expected present value of the firm at time 0, resulting from production and investment decisions over  $t=1, 2, \dots, M$ , is

$$(11) \quad V_0 = \sum_{t=1}^{t=M} (1+r)^{-t}(p'Q_t - w'L_t - q'I_t)$$

This can be restated in terms of  $K_t$  by substituting in equation (2) and rearrang-

ing terms so that<sup>5</sup>

$$(12) \quad V_0 = \sum_{t=1}^{t=M} (1+r)^{-t}(p'Q_t - w'L_t - u'K_t)$$

where  $u' = q'(r + \delta)$ . The expression  $u'$  is the after tax "rental price" of capital.<sup>6</sup> Equation (12) and the production function equation (1) indicate how tax parameters  $t_b$ ,  $t_v$ ,  $t_i$ ,  $t_s$ , and  $v$ ; market prices  $p$ ,  $q$ ,  $w$ , and  $r$ ; and the nature of the production process determine the increment in present value of the firm resulting from different choices of  $K_t$  and  $L_t$  over the firm's planning horizon.

## II. Industry Investment Models and Empirical Results

Up to this point we have developed a model relating prices and tax and monetary parameters to the present value of the *firm*. This section derives *industry* investment models based on equation (12) and tests the resulting investment equations using annual data, 1950-65, for the manufacturing sector in France and eight industry subgroups of manufacturing.<sup>7</sup> Ordinary least squares regression techniques are employed to estimate the relative explanatory power of the models and of different individual variables. The sensitivity of empirical results to changes in lag structure and the inclusion of time trend is also included in the analysis.

As a first approximation assume the forces operating at the level of the firm under perfect competition operate in the same fashion at the industry level. Thus a firm maximizing its present value, equa-

<sup>5</sup> This expression does not include the fixed costs,  $q'K_1$ , arising from the existing capital stock, and neglects the ending term  $(1+r)^{-M}q'K_{M+1}$ , assuming large  $M$ .

<sup>6</sup> For a discussion of the rental price of capital, see Jorgenson, Hall and Jorgenson (1967) and Coen (1968).

<sup>7</sup> Detailed data descriptions and source information are available from the author upon request. Most of the data used were from publications of the Institut National de la Statistique et des Études Économiques, Paris, France.

tion (12), would choose  $K_t$  and  $L_t$  to satisfy the traditional marginal productivity conditions

$$(13a) \quad \frac{\partial Q_t}{\partial K_t} = \left(\frac{u}{p}\right)'$$

$$(13b) \quad \frac{\partial Q_t}{\partial L_t} = \left(\frac{w}{p}\right)'$$

Approximating the production function equation (1) as Cobb-Douglas

$$(14) \quad Q_t = A e^{at} K_t^\alpha L_t^\beta$$

with  $0 \leq \alpha, \beta \leq 1$ ,  $g \geq 0$ , and  $A$  constant, the level of  $K_t$  satisfying equations (13) and (14) is (in logs)

$$(15) \quad \ln K_t^* = a_0 + a_1 \ln \left(\frac{u}{p}\right)' + a_2 \ln \left(\frac{w}{p}\right)' + a_3 t$$

where

$$a_0 = \left(\frac{\beta - 1}{\gamma} \ln \alpha - \frac{\beta}{\gamma} \ln \beta - \frac{1}{\gamma} \ln A\right)$$

$$a_1 = \left(\frac{1 - \beta}{\gamma}\right)$$

$$a_2 = \left(\frac{\beta}{\gamma}\right)$$

$$a_3 = \left(\frac{-g}{\gamma}\right)$$

$$\gamma = \alpha + \beta - 1$$

If we represent actual investment behavior as<sup>8</sup>

$$(16) \quad \left(\frac{K_{t+1}}{K_t}\right) = \left(\frac{K_t^*}{K_t}\right)^b \text{ with } 0 \leq b \leq 1,$$

then the resulting testable investment equation is

$$(17) \quad \begin{aligned} & \ln (K_{t+1}/K_t) \\ &= ba_0 + ba_1 \ln \left(\frac{u}{p}\right)' + ba_2 \ln \left(\frac{w}{p}\right)' \\ &+ ba_3 t - b \ln K_t \end{aligned}$$

If  $\alpha + \beta < 1$  (diminishing returns to scale) then  $ba_1, ba_2 \leq 0$  and  $ba_3 \geq 0$ .

Before interpreting the empirical results presented in this section it should be noted that the multiple correlation coefficient adjusted for degrees of freedom ( $R^2 \text{adj.}$ ) can be used to compare equations but does not represent a meaningful measure of explanatory power for individual equations. This is because adding  $\ln K_t$  to both sides, which changes the dependent variable to  $\ln K_{t+1}$ , provides identical coefficient estimates for all variables except  $\ln K_t$  but raises all  $R^2 \text{adj.}$  Furthermore, since the Durbin-Watson measure of serial correlation is biased towards levels indicating no serial correlation in autoregressive models, it is presented to test where serial correlation exists and not to indicate its absence.

Table 1 presents the regression results for equation (17) with one or two lagged capital stock variables. If the inclusion of a second lagged capital stock variable increased the  $R^2 \text{adj.}$ , then these regression results were reported. Expected prices were assumed to be based on the average price levels prevailing in the preceding year ( $t-1$ ) and capital stock estimates refer to the beginning of period  $t$  and  $t-1$ .

In general, the empirical results of Table 1 provide little evidence of an important role for relative prices in this model of industry investment. While the sign pattern on  $\ln (u/p)'$  is as anticipated, the sign pattern on  $\ln (w/p)'$  is mixed and most coefficients on both price variables are not statistically significant. Furthermore, if these results are used to compute the price elasticities  $a_1$  and  $a_2$  from equation (15), in almost all cases the sum of the price elasticities exceeds  $-1.0$

<sup>8</sup> Hickman employs this adjustment process.

TABLE 1—REGRESSION RESULTS: INVESTMENT EQUATION (17)

Industry	Constant	$\ln (u/p)_{t-1}$	$\ln (w/p)_{t-1}$	$\ln K_t$	$\ln K_{t-1}$	t	R <sup>2</sup> adj.	d
Total Private Manufacturing	4.000 (1.572)	-.1516 (.063)	-.2497 (.134)	-.3966 (.152)		.0333 (.012)	.4057	1.20
Iron and Steel	3.665 (1.149)	-.0986 (.086)	.0262 (.100)	.4714 (.192)	-.9173 (.215)	.0224 (.007)	.6490	2.00
Nonferrous Metals	2.248 (1.031)	.0172 (.149)	-.0154 (.170)	.3224 (.249)	-.6605 (.244)	.0147 (.010)	.2677	2.56
Chemicals and Rubber	1.652 (1.022)	-.00731 (.060)	-.0219 (.120)	.4305 (.224)	-.6293 (.230)	.0146 (.013)	.3102	2.32
Construction Materials and Glass	-.8060 (.446)	-.0472 (.050)	.2270 (.099)	.0953 (.062)		-.00999 (.006)	.8647	2.49
Mechanical and Elec. Equipment	8.803 (1.961)	-.0664 (.177)	-.3519 (.367)	-.4784 (.227)	-.4793 (.190)	.0725 (.022)	.6258	1.25
Textiles, Clothing, and Leather	2.296 (.578)	-.0933 (.033)	-.0205 (.052)	.2992 (.282)	-.5748 (.286)	.00963 (.004)	.9116	2.15
Wood, Paper, and Misc.	.7359 (.549)	-.0523 (.042)	.03319 (.091)	-.0957 (.064)		.00506 (.006)	.8586	1.69
Food and Beverages	3.598 (1.502)	-.1002 (.033)	.0251 (.058)	-.0756 (.176)	-.3541 (.198)	.0150 (.007)	.5432	2.78

Notes: Standard errors of coefficients are in parentheses; R<sup>2</sup>adj. = Multiple correlation coefficient adjusted for degrees of freedom; d = Durbin-Watson statistic

implying  $\alpha + \beta \leq 0$ , i.e., near zero or negative marginal productivities of capital and labor. The sign pattern on t and on  $\ln K_t$  (or the sum of coefficients on  $\ln K_t$  and  $\ln K_{t-1}$ ) is as anticipated and these coefficients are more frequently significant than those on the price variables. But overall one would have to conclude the empirical results from Table 1 are essentially negative.

As a second and more reasonable approximation of industry investment behavior consistent with many accelerator studies of U.S. investment,<sup>9</sup> we assume

<sup>9</sup> See, for example, Eisner (1960, 1962), Hickman, and Coen (1968). Jorgenson's basic investment equation is a special case of this model which essentially treats labor as well as output as exogenous. Under these conditions Jorgenson uses only the marginal productivity condition for capital, equation (13a), to derive his expression for desired capital stock,  $K^* = \alpha p Q / u$ .

that firms treat output (or sales) as exogenous and adjust capital and labor to minimize the present value of total factor costs. Maximizing equation (12) subject to a fixed output level results in the first-order condition

$$(18) \quad \frac{\partial Q_t}{\partial L_t} / \frac{\partial Q_t}{\partial K_t} = \left( \frac{w}{u} \right)'$$

Substituting in equation (14) and solving for  $\ln K_t^*$  for given Q we find<sup>10</sup>

$$(19) \quad \ln K_t^* = a_4 + a_5 \ln Q + a_6 \ln \left( \frac{w}{u} \right)' + a_7 t$$

<sup>10</sup> This proof is presented in Marc Nerlove, pp. 106-07.

where

$$a_4 = \alpha [A \alpha^\alpha \beta^\beta]^{-\frac{1}{\alpha+\beta}}$$

$$a_5 = \frac{1}{\alpha + \beta}$$

$$a_6 = \frac{\beta}{\alpha + \beta}$$

$$a_7 = \frac{g}{\alpha + \beta}$$

Employing equation (16) the resulting investment equation is

$$\begin{aligned} \ln (K_{t+1}/K_t) \\ (20) \quad &= ba_4 + ba_5 \ln Q + ba_6 \ln \left( \frac{w}{u} \right)' \\ &+ ba_7 t - b \ln K_t \end{aligned}$$

with  $ba_5$ ,  $ba_6$ , and  $ba_7 > 0$ .

Table 2 presents the regression results for equation (20), with expected output level measured as the average output level of the previous year.<sup>11</sup> These results are not substantially better than those of Table 1. There is an improved fit for the primary metal industries, little change in other industries, and a poorer fit for aggregate private manufacturing. While the signs of the coefficients on  $\ln Q_{t-1}$ ,  $\ln (w/u)_{t-1}$ , and the capital stock variables are consistent with theoretical expectations, most coefficients are not statistically significant. The time trend variable has lost its importance, suggesting it may have served as a proxy for an expected output variable. This conclusion is supported in

<sup>11</sup> The time variable was dropped when its inclusion lowered the  $R^2$ adj. When included, however, its coefficient was positive.

TABLE 2—REGRESSION RESULTS: INVESTMENT EQUATION (20)

Industry	Constant	$\ln Q_{t-1}$	$\ln (w/u)_{t-1}$	$\ln K_t$	$\ln K_{t-1}$	t	$R^2$ adj.	d
Total Private Manufacturing	.7192 (1.002)	.0377 (.120)	.0849 (.071)	-.0911 (.136)			.0771	.79
Iron and Steel	1.394 (.319)	.2413 (.048)	.0495 (.053)	.5528 (.134)	-.8406 (.136)		.8040	1.45
Nonferrous Metals	1.501 (.609)	.1889 (.097)	-.0168 (.135)	.1899 (.266)	-.5274 (.239)		.3490	2.35
Chemicals and Rubber	.9739 (.655)	.1423 (.081)	.00498 (.049)	.2028 (.271)	-.3851 (.219)		.3521	2.16
Construction Materials and Glass	-.3848 (.345)	.02545 (.062)	.0825 (.048)	.0200 (.073)			.8414	2.22
Mechanical and Elec. Equipment	7.819 (2.284)	.3670 (.332)	.0488 (.166)	-.6230 (.212)	-.3854 (.181)	.0306 (.028)	.6335	1.46
Textiles Clothing, and Leather	.9227 (.381)	.1048 (.038)	.0665 (.033)	-.1677 (.045)			.8774	1.46
Wood, Paper, and Misc.	.5519 (.304)	.1152 (.064)	.0345 (.034)	-.1276 (.055)			.8843	1.87
Food and Beverages	3.584 (1.718)	-.0416 (.111)	.1179 (.033)	-.0390 (.195)	-.3702 (.220)	.01227 (.008)	.4533	2.50

Notes: See Table 1

later tests which include a demand shift variable.

As an additional check on the reliability of the coefficient estimates in the investment equations, these results were used to estimate the production parameters  $\alpha$  and  $\beta$ . The relations between the investment equation coefficients and the underlying production parameters can be seen by examining equations (19) and (20). In all cases the sum of  $\alpha$  and  $\beta$  exceeded 1.0 and in many cases the individual estimates of  $\alpha$  and  $\beta$  were far from reasonable values. Again we must conclude that our empirical results are essentially negative.

In the two industry investment models tested above, we have applied models of the behavior of an individual firm directly to the industry level without modifying market conditions to reflect the effects of aggregation. In particular we have treated either price or quantity as exogenous influences when these should more appropriately be treated as endogenous at the industry level if the industry is relatively competitive. As a final extension of the neoclassical approach, one that has received little attention in studies of U.S. investment,<sup>12</sup> we will add the following output demand equation

$$(21) \quad Q = Bp^\eta Y^\xi$$

where  $p$  is output price,  $Y$  is a demand shift variable, total national product, and both are deflated by a national product deflator. The terms  $\eta$  and  $\xi$  represent price and "income" elasticities. Solving equations (21), (13), and (14) for  $\ln K_t^*$  we have

$$(22) \quad \ln K_t^* = a_8 + a_9 \ln w'' + a_{10} \ln u'' \\ + a_{11} \ln Y_t + a_{12} t$$

where  $w''$  and  $u''$  are  $w'$  and  $u'$  divided by  $(1-t_v)(1-t_b)$  and a national product deflator, and where

$$a_8 = \frac{1}{\gamma} \{ \ln B - (1 + \eta) \ln A - \beta(1 + \eta) \ln B \\ + [\beta(1 + \eta) - \eta] \ln \alpha \}$$

$$a_9 = \frac{\beta(1 + \eta)}{\gamma}$$

$$a_{10} = \frac{\eta - \beta(1 + \eta)}{\gamma}$$

$$a_{11} = \frac{\xi}{\gamma}$$

$$a_{12} = - \frac{(1 + \eta)g}{\gamma}$$

$$\gamma = (\alpha + \beta)(1 + \eta) - \eta$$

Finally, from equation (16) industry investment becomes

$$(23) \quad \ln (K_{t+1}/K_t) \\ = ba_8 + ba_9 \ln w'' + ba_{10} \ln u'' \\ + ba_{11} \ln Y + ba_{12} t - b \ln K_t$$

In general we should expect  $ba_{10} < 0$  and  $ba_{11} > 0$  with the sign of  $ba_9$  and  $ba_{12}$  sensitive to the magnitude of the price elasticity of demand.

Table 3 presents the results for equation (23) measuring all expected values with last period's average values. These results, which are still not very impressive, explain investment in chemicals and rubber, mechanical and electrical products, and food and beverages better than previous models; do not explain primary metal industry investment as well as equation (20), and represent little change in overall explanatory power for the other industries. The sign pattern on  $\ln u''$ ,  $\ln Y$  and the capital stock variables is generally as predicted, the time variable loses its importance relative to Table 1 results, and the sign pattern on  $\ln w''$  is mixed. Most coefficients are not statistically significant.

If we assume constant returns to scale the results from Table 3 can be used to

<sup>12</sup> This approach was used by Robert Lucas.

TABLE 3—REGRESSION RESULTS: INVESTMENT EQUATION (23)

Industry	Constant	$\ln w''_{t-1}$	$\ln w'_{t-1}$	$\ln F_{t-1}$	$\ln K_t$	$\ln K_{t-1}$	t	R <sup>2</sup> adj.	d
Total Private Manufacturing	6.824 (5.076)	-.3394 (.259)	-.1208 (.063)	.3502 (.679)	-.4419 (.352)		.0175 (.013)	.325	1.60
Iron and Steel	3.993 (2.474)	.1422 (.206)	-.0822 (.082)	.4353 (.356)	.5100 (.228)	-1.006 (.220)		.556	1.56
Nonferrous Metals	2.0255 (3.067)	.0296 (.297)	.0002 (.147)	.2652 (.406)	.4189 (.260)	-.7336 (.284)		.2345	2.54
Chemicals and Rubber	4.399 (2.521)	-.2316 (.137)	-.0581 (.049)	.0526 (.357)	.0203 (.307)	-.3914 (.212)	.0309 (.010)	.5365	2.57
Construction Materials and Glass	-1.679 (1.707)	.1906 (.149)	-.0733 (.055)	-.1605 (.230)	.1241 (.125)			.8325	2.40
Mechanical and Electrical Equipment	16.798 (1.231)	-.6295 (.119)	.0330 (.0750)	2.261 (.189)	-1.218 (.113)	-.1942 (.091)		.9387	2.55
Textiles, Clothing, and Leather	3.506 (1.316)	-.1409 (.121)	-.1120 (.046)	.2015 (.096)	.2912 (.376)	-.5617 (.382)		.8945	2.13
Wood, Paper, and Misc.	2.027 (2.015)	-.0139 (.108)	-.0562 (.044)	.2325 (.237)	-.2002 (.172)			.8580	1.93
Food and Beverages	5.432 (1.402)	-.0428 (.073)	-.0698 (.029)	.4840 (.139)	-.4040 (.203)	-.1507 (.138)		.6966	2.62

Notes: See Table 1.

provide estimates of the income and price elasticities of demand,  $\xi$  and  $\eta$ .<sup>13</sup> The estimates for income elasticities range from .74 (textiles) to 1.60 (mechanical and electrical equipment). Excluding the metals industries, the price elasticities range from -.20 (food) to -1.04 (total manufacturing). Since these results from extended product market specification are reasonable, they suggest the inadequacies in the models have their roots in misspecification of the factor market influences on investment. We will comment further on this below.

### III. Concluding Remarks

For analysis of French investment experience the theoretical tools of neoclassical analysis, such as the concepts of net present value and capital rental prices, prove to be very useful. This paper demonstrates how these concepts can be used to incorporate major product and factor market influences, including the complexi-

ties of the French tax system, into an expression for the firm's net present value. When coupled with assumptions about the objectives of the firm, the nature of the production function, and the capital stock adjustment process, this approach provides a set of theoretically consistent investment demand equations.

While the neoclassical approach and the resulting investment demand equations may appeal to our belief in clearly specifying the theoretical basis of demand equations, the empirical results of Section II raise serious questions about the adequacy of these investment models in explaining actual French investment behavior. Although the empirical results improved somewhat with refinement of the industry investment models, taken as a whole they must be viewed as essentially negative.

These results can be explained in several ways. If we assume the single equation neoclassical investment demand approach is valid, the negative results may be due to the production function choice, the as-

<sup>13</sup> If  $\alpha + \beta = 1$ , then from equation (22) we see that  $a_{11} = \xi$  and  $a_9 + a_{10} = \eta$ .

sumed capital stock adjustment process, the way expected levels of variables are estimated or inadequacies in the data or the final forms of the equations used for testing. All of these factors may account for empirical problems and have been raised frequently in criticisms of U.S. investment studies.

While the above factors may have affected my empirical results, reliance on a single equation investment demand approach may well represent a more important problem. Although my model was extended to incorporate more complex product demand conditions, all equations ignored what may have been critical changes in the conditions influencing the supply of investment goods and investment funds<sup>14</sup> over the period 1950-65. This may explain why the capital rental price variable, which encompasses the price of both capital goods and capital funds, performs so poorly in *French* compared to U.S. investment equations. If the empirical results are picking up important unspecified supply effects, the single equation investment demand approach characteristic of studies of U.S. investment behavior may not be readily applied to studies of investment behavior in other countries. At least we have some evidence that this conventional approach may be inadequate to explain investment behavior in France.

#### REFERENCES

- C. W. Bischoff, "Hypothesis Testing and the Demand for Capital Goods," *Rev. Econ. Statist.*, Aug. 1969, 51, 354-68.
- , "Lags in Fiscal and Monetary Impacts on Investment in Producers' Durable Equipment," in G. Fromm, ed., *Tax Incentives and Capital Spending*, Washington 1970.
- R. M. Coen, "Effects of Tax Policy on Investment in Manufacturing," *Amer. Econ. Rev. Proc.*, May 1968, 58, 200-11.
- , "Tax Policy and Investment Behavior: Comment," *Amer. Econ. Rev.*, June 1969, 59, 370-79.
- R. Eisner, "A Distributed Lag Investment Function," *Econometrica*, Jan. 1960, 28, 1-29.
- , "Investment Plans and Realizations," *Amer. Econ. Rev. Proc.*, May 1962, 52, 190-203.
- , "Tax Policy and Investment Behavior: Comment," *Amer. Econ. Rev.*, June 1969, 59, 379-88.
- and R. H. Strotz, "Determinants of Business Investment," in Commission on Money and Credit, *Impacts of Monetary Policy*, Englewood Cliffs 1963.
- , and M. I. Nadiri, "Investment Behavior and the Neo-Classical Theory," *Rev. Econ. Statist.*, Aug. 1968, 50, 369-82.
- and ———, "Neo-Classical Theory of Investment Behavior: A Comment," *Rev. Econ. Statist.*, May 1970, 52, 216-22.
- M. K. Evans, *An Econometric Model of the French Economy*, OECD, Paris, Mar. 1969.
- T. Haavelmo, *A Study in the Theory of Investment*, Chicago 1960.
- R. E. Hall and D. W. Jorgenson, "Tax Policy and Investment Behavior," *Amer. Econ. Rev.*, June 1967, 57, 391-414.
- and ———, "Tax Policy and Investment Behavior: Reply and Further Results," *Amer. Econ. Rev.*, June 1969, 59, 388-401.
- B. G. Hickman, *Investment Demand and U.S. Economic Growth*, Washington 1964.
- D. W. Jorgenson, "Capital Theory and Investment Behavior," *Amer. Econ. Rev. Proc.*, May 1963, 53, 247-59.
- and C. D. Siebert, "A Comparison of Alternative Theories of Corporate Investment Behavior," *Amer. Econ. Rev.*, Sept. 1968, 58, 681-712.
- and J. A. Stephenson, (1967a) "The Time Structure of Investment Behavior in United States Manufacturing, 1947-1960," *Rev. Econ. Statist.*, Feb. 1967, 49, 16-27.
- and ———, (1967b) "Investment Behavior in U.S. Manufacturing," *Econometrica*, Apr. 1967, 35, 169-220.

<sup>14</sup> Michael Evans, in a study of total French fixed business investment, 1952-65, presents evidence in support of an important role for financing costs in the investment decision.

- and ———, "Issues in the Development of the Neo-Classical Theory of Investment Behavior," *Rev. Econ. Statist.*, Aug. 1969, 51, 346-53.
- R. E. Lucas, Jr., "Substitution between Labor and Capital in U.S., Manufacturing, 1929-1958," unpublished doctoral dissertation, Univ. Chicago 1963.
- M. Nerlove, *Estimation and Identification of Cobb-Douglas Production Functions*, Chicago 1965.
- J. C. R. Rowley, "Investment Functions: Which Production Function?," *Amer. Econ. Rev.*, Dec. 1970, 60, 1008-12.
- R. Schramm, "The Influence of Relative Prices, Production Conditions and Adjustment Costs on Investment Behavior," *Rev. Econ. Stud.*, July 1970, 37, 361-76.
- C. K. Sullivan, *The Tax on Value-Added*, New York 1965.
- Institut National de la Statistique et des Études Économique (INSEE), *Études et Conjonctures*, various issues.
- , *Annuaire Statistiques de la France*, various issues.
- , *Bulletin Mensuel de Statistique*, various issues.

CHIEF  
 1. 1. 1. 1.  
 29111  
 50  
 11  
 1. 1. 1.

# Anticipatory and Objective Models of Durable Goods Demand

F. THOMAS JUSTER AND PAUL WACHTEL\*

The propensity of U.S. households to acquire tangible assets like automobiles and household appliances at varying rates over time still remains as one of the less well-understood and predictable aspects of economic behavior. In part, the explanation may be that consumption research has tended to focus on real consumption (use) flows and not on consumer expenditure and investment decisions. A second reason for the present unsatisfactory state of knowledge, and for our inability to predict near-term consumer behavior with reasonable accuracy, may lie in the failure of most model builders to explore seriously the use of data on consumer anticipations as an adjunct to the more traditional information on asset stocks and income flows that models generally tend to emphasize. This paper examines that possibility. First, we develop a nonanticipatory (objective) model of consumer durable goods demand, then we contrast the performance of this objective model with one based largely on the use of survey measures of consumer anticipations, and in the last section we examine the characteristics of an optimal model which combines both types of information.

A commonly used framework for anal-

ysis of consumer behavior, the stock adjustment model, views households as attempting to adjust actual to desired stocks of assets. Within this framework, survey measures of consumer purchase expectations can be interpreted as a subjective estimate of the difference between actual and desired stock, with reported purchase expectations reflecting the speed of the adjustment process as well as the underlying determinants of desired stock. And survey measures of consumer attitudes (optimism, pessimism) might be interpreted as one of the arguments in the desired stock function.

Demand models based on survey measures of consumer anticipations can be contrasted with models that exclude them and rely wholly on objective variables like income, price, and the stock of durables, as well as with joint models that incorporate both types of variables. Although a number of studies have explored this question, none has done so thoroughly or systematically. Typically, they have focused on examining the usefulness of anticipatory variables in a more or less *ad hoc* context; that is, objective variables have been introduced into demand models along with anticipations in order to determine whether the anticipations were significantly associated with purchases after accounting for the influence of income, etc.<sup>1</sup>

Studies concerned with the specification

\* National Bureau of Economic Research and City College of the City University of New York and National Bureau of Economic Research, respectively. This paper is part of a larger study being carried out by the National Bureau of Economic Research in cooperation with the U.S. Bureau of the Census. The study is being financed in part with the aid of a grant from the National Science Foundation. We wish to express our appreciation to Robert Eisner, Paul Taubman, and David Kresge who read the manuscript as well as to Saul Hymans who contributed valuable suggestions.

<sup>1</sup> The work by Eva Mueller, E. Scott Maynes, and Juster (1969a) fits into this general framework, in that the main focus is on the performance of anticipatory variables in a demand model. All pay only incidental attention to the structure of an objective model.

of an objective model have not ordinarily shown much interest in the potential uses of anticipatory data. This is in part because such models have been concerned with the role of basic economic variables like income and prices in the explanation of purchase behavior, and not with the possible forecasting uses of the model.<sup>2</sup> And even where forecasting uses have been an important element in determining the structure of the model, for example, in the consumer durables equations of econometric models, only rarely have the model builders attempted to incorporate anticipatory data.<sup>3</sup>

For the purpose of explaining consumer behavior, anticipatory variables like intentions or attitudes tend to muddy the coefficients of objective variables like income and prices because the two sets of variables reflect roughly the same economic phenomena. Thus to estimate the influence of, for example, income on purchases in a model that includes both income and buying intentions, it is necessary to estimate the influence of income on intentions, then add this to the measured influence of income. In models designed for forecasting, the anticipatory variables are often difficult to use because they tend to cover a limited time span and often have to be extensively processed before they can be used effectively. Moreover, simulation of the model requires that future values of the anticipatory variables be predicted. If they could be accurately predicted, one would not need them in the first place, and if the predictions are poor, the simulation is unsatisfactory.<sup>4</sup>

<sup>2</sup> For examples of demand studies of this type see Gregory Chow, Daniel Suits, and Michael Hamburger.

<sup>3</sup> The current versions of the Brookings, Wharton, *FMP*, and *OBE* models use some form of the stock adjustment process, and none contain anticipatory variables. Earlier versions of the Brookings and Wharton models included the Index of Consumer Sentiment as an explanatory variable.

<sup>4</sup> A recent paper by Saul Hymans uses the Index of

## I. The Objective Demand Model

Durable goods yield utility to consumers in the form of a flow of services which continues until the product is fully depreciated. The analysis of demand for consumer durables therefore focuses on the demand for durable goods stock, and only indirectly examines purchases. In this section, we develop a model that relates several aspects of purchase behavior to stock demand. The model makes provision for the lagged adjustment of the stock of durables to changes in the equilibrium level of stocks, for the expectational basis of stock demand, and for the distinction between transitory and permanent influences on demand.

### *Specification of the Model*

In general terms, the model views consumers as having a "target" or "desired" value of durables stocks to which they adjust gradually.<sup>5</sup> Net investment is viewed as having a "permanent" or "planned" component and also an "unforeseen" or "transitory" component.<sup>6</sup> The permanent component depends on long-run expectations and average adjustment lags, while the transitory com-

---

Consumer Sentiment in a model designed to be simulated, with an auxiliary prediction equation for the Index itself.

<sup>5</sup> Adjustments are not made instantaneously partly because of decision and purchasing lags, partly because the level of desired stock represents a target demand about which there exists some uncertainty, and partly because of transactions costs. Household investment decisions are sensitive to uncertainty because resale markets are imperfect; a decision to invest represents a commitment to consume a certain level of services well into the future. Increasing marginal costs of investment are usually cited in the capital investment literature as the source of adjustment lags.

<sup>6</sup> In our model, the distinction between permanent and transitory investment is the length of the planning horizon that precedes the investment decision. Thus, transitory investment may come from an unexpected but permanent income change which alters the rate of consumption and therefore the level of durable stock held.

ponent represents the immediate reaction to unexpected income flows. The transitory component accounts for the volatile behavior of investment because unforeseen economic phenomena alter the time pattern of stock adjustments.

The partial adjustment model is applied to the planned component of net durables investment,  $\Delta S^P$ , as in (1) where  $\beta$  represents the average speed at which households move to desired stock levels;  $S^*$ , the level of desired stock

$$(1) \quad \Delta S^P = \beta(S^* - S_{-1})$$

is a target set by the household contingent upon its expectations about economic conditions. Given expectations, there is some level of stocks the household would like to hold, and it plans to close some proportion of the gap between existing and desired stocks during the current period.

Desired stock is a function of expected values of a set of economic variables denoted by  $Z$ . The specification of variables in the  $Z$  function is discussed below; the expectation is shown in (2).

$$(2) \quad S^* = Z^e$$

The  $Z$  function is taken to be linear, and expectations are generated by the uniform application of the adaptive expectations hypothesis to all variables in  $Z$ . The adaptive expectation model for the formation of expectations by the household is given in (3).

$$(3) \quad Z^e - Z_{-1}^e = \rho(Z - Z_{-1}^e)$$

The specification shown here is in the form of a discrete approximation to a continuous revision procedure, rather than a discrete version of the model.<sup>7</sup> This difference de-

<sup>7</sup> The discrete model, in contrast, states that the current expectation differs from the previous expectation by some proportion of the error made in the last period. The correct specification of this model is the continuous form (3'), as expectations are being continually revised.

$$(3') \quad \frac{dZ^e}{dt} = \rho(Z - Z^e)$$

termines whether the current or lagged value of  $Z$  appears in the model. The interpretation of (3) is that the change in expectations is proportional to the difference between current experience and the previously formed expectation.<sup>8</sup>

The last element of the model is the transitory investment component (4), a function,  $T$ , of transitory variables specified below.

$$(4) \quad \Delta S^T = T$$

Equation (5) defines net investment as the sum of its transitory  $\Delta S^T$ , and permanent,  $\Delta S^P$ , components.

$$(5) \quad \Delta S = \Delta S^P + \Delta S^T$$

The reduced form of the model given by (1)–(5) is a second-order distributed lag which describes the effect on durables stocks of the change in an economic variable in the  $Z$  and  $T$  functions.<sup>9</sup> The model, which results from the convolution of two first-order lag models, is shown in dif-

The approximation to (3') is (3), which is the form used above. Its only drawback is that it is not an *ex ante* explanation, since it requires current observations to explain current expectations. When a pure forecast form is required the discrete error revision version (3'') of the model can be substituted:

$$(3'') \quad Z^e = Z_{-1}^e + \rho(Z_{-1} - Z_{-1}^e)$$

<sup>8</sup> The symmetry of the partial adjustment and adaptive expectations first-order lag models has been discussed by Roger Waud.

<sup>9</sup> The reduced form is derived by writing the model in terms of lag operators. We can rewrite the identity in (5) in terms of stock, then substitute (1) and (4), all expressed with the lag operator  $L$  to yield:

$$S - LS = \beta S^* - \beta LS + T$$

Using (2), substitute for  $S^*$  and solve for  $S$ :

$$S = \frac{1}{1 - (1 - \beta)L} [\beta Z^e + T]$$

Similarly, (3) can be solved for  $Z^e$  as a weighted sum of lagged values of  $Z$ , and substituted above to yield the reduced form:

$$S = \left[ \frac{\beta}{1 - (1 - \beta)L} \right] \left[ \frac{\rho}{1 - (1 - \rho)L} \right] Z + \frac{T}{1 - (1 - \beta)L}$$

ference equation form as (6).

$$\begin{aligned} S &= \rho\beta Z + [(1 - \beta) + (1 - \rho)]S_{-1} \\ (6) \quad & - (1 - \beta)(1 - \rho)S_{-2} + T \\ & - (1 - \rho)T_{-1} \end{aligned}$$

The lag parameters  $\rho$  and  $\beta$  are the coefficient of expectations and the speed of adjustment, respectively; however, the full model involves the two lag processes concurrently and individual estimates have no interpretation even when identified. If expectations are formed instantaneously,  $\rho = 1$ , and the model reduces to a first-order lag scheme. If adjustments are made instantaneously,  $\beta = 1$ , and the model reduces to a similar first-order scheme. Thus a first-order model can be derived from either lag model, each being a special case of the complete model. A first-order model would be suggested if the coefficient on  $S_{-2}$  is insignificant; otherwise misspecification would result in sizeable biases. Waud's Monte Carlo study indicates that a partial adjustment model that ignores the adaptive formation of expectations produces a downward bias in the speed of adjustment and an exaggeration of the standard errors.

The model actually estimated has net or gross investment rather than stock as the dependent variable, and is obtained by subtracting  $S_{-1}$  from both sides of (6) and rearranging terms to yield (7). This is the full objective model, which we call *AET* (partial adjustment-adaptive expectations-transitory change).

$$\begin{aligned} \Delta S &= \rho\beta Z - \rho\beta S_{-1} \\ (7) \quad & + (1 - \rho)(1 - \beta)\Delta S_{-1} \\ & + T - (1 - \rho)T_{-1} \end{aligned}$$

A test of this version of the reduced form is that the current and lagged transitory terms are specified to be of opposite sign with the lagged term smaller in absolute value because  $(1 - \rho) < 1$ .

The model can be readily translated from net to gross investment by using the identity  $G = \delta S_{-1} + \Delta S$ , where  $G$  is purchases and  $\delta$  is the depreciation rate; this version is shown as equation (7.1).<sup>10</sup>

$$\begin{aligned} G &= \rho\beta Z + (\delta - \rho\beta)S_{-1} \\ (7.1) \quad & + (1 - \rho)(1 - \beta)\Delta S_{-1} \\ & + T - (1 - \rho)T_{-1} \end{aligned}$$

Simplified versions of the model are also tested. The reduced form (8) ignores the distinction between planned and transitory components of net investment.

$$\begin{aligned} \Delta S &= \rho\beta Z - \rho\beta S_{-1} \\ (8) \quad & + (1 - \rho)(1 - \beta)\Delta S_{-1} \end{aligned}$$

This is the full model without the transitory change component (*AE*). It can be estimated with permanent, current or both permanent and transitory income as elements of  $Z$ .

A first-order adjustment model, derived by setting the coefficient of expectations equal to unity, is also tested in (9).

$$(9) \quad \Delta S = \beta Z - \beta S_{-1} + T$$

This is the partial adjustment-transitory change model (*AT*). In gross investment form (9.1) this is the model most com-

<sup>10</sup> An alternative formulation of the gross investment model specifies that a partial adjustment to depreciated stock determines planned purchases.

$$(1') \quad G^p = \beta(S^* - (1 - \delta)S_{-1})$$

The reduced form of the model specified by (1'), (2), (3), (4'), and (5') is given by (8') which is identical to (8) except that the coefficients have a different interpretation.

$$(4') \quad G^T = T$$

$$(5') \quad G = G^p + G^T$$

$$\begin{aligned} (8') \quad G &= \beta\rho Z + [(1 - \rho)\delta - \beta\rho(1 - \delta)]S_{-1} \\ & + (1 - \rho)(1 - \delta)\Delta S_{-1} + T - (1 - \rho)T_{-1} \end{aligned}$$

This form would be preferred if supply restrictions or an absolute decline in wealth led the household sector to delay replacement demand. There is no evidence that this occurs in the sample period. Net investment in total durables is never negative and the automobile component is less than zero in only five quarters of the twenty-year period examined.

monly found in the econometric literature.

$$(9.1) \quad G = \beta Z + (\delta - \beta)S_{-1} + T$$

This model, without a transitory term, was introduced by Suits, Chow, and others. Richard Stone and D. A. Rowe and Hamburger make use of specific depreciation assumptions to derive a reduced form in lagged purchases without any explicit estimate of the total stock.

### *Empirical Estimation of the Model*

The models outlined in the preceding section are estimated for the period 1949 through 1967, using quarterly data. Equations with both net investment  $N$  and gross investment  $G$  as dependent variables are examined; results are shown for total durables, and for automobile and household durables separately (denoted by  $D$ ,  $C$ , and  $H$  subscripts, respectively). All variables representing value aggregates are deflated per household magnitudes (1958 prices).<sup>11</sup>

The set  $Z$  is composed of the price and income variables that determine the desired stock target. The relevant price variables are all relative prices, the series being the respective implicit price deflators,  $P$ , relative to the deflator for total personal consumption expenditure,  $Q$ .

For the automobiles and total durables models a measure of credit availability or cost is also used. The measure we use,  $M$ , the maturity on instalment credit contracts, has often been found to have a strong influence on purchases. Contract maturity and unit price determine the amount of the monthly instalment payment, which is an important factor in determining the number of credit purchases. The maturity variable also reflects a price effect via its relation to the true marginal

borrowing cost for consumers subject to credit rationing. Results using the pure price of credit, the interest rate, as an alternative credit variable are discussed in an Appendix to the *NBER* reprint of this paper.<sup>12</sup>

The uniform application of adaptive expectations may be unwarranted for the income variable. Therefore, permanent and transitory income variables,  $Y^*$  and  $Y^T$ , were explicitly estimated.<sup>13</sup> The models are estimated with permanent or current disposable income,  $Y$ , as alternative income variables in the  $Z$  function.

All regressions are estimated with a set of dummy variables that represent abnormal supply conditions. Panic buying during the Korean War, which resulted from fears of shortages, is treated in this way, as are the three strikes which affected the automobile market. The Korean War dummies  $KD$  are designed to minimize residuals in 1950-III, IV, and 1951-I. A uniform strike and poststrike recovery dummy  $SD$  is used for 1952-III, 1959-IV, and 1964-IV. In the second-order lag models, abnormal supply conditions affect not only the dependent variable but also bias the coefficient of the lagged dependent variable specified by the model;

<sup>12</sup> Consumer purchases are not ordinarily thought to be sensitive to changes in interest rates per se, which often are not adequately reflective of conditions in credit markets. In terms of financial flows, an increased interest cost is readily balanced out by a longer maturity. This can be seen by looking at the value of a loan,  $V = \pi(1 - (1+i)^{-M})/i$ , where  $\pi$  is the monthly payment,  $i$  the loan rate and  $M$  the maturity. The elasticity of  $V$  with respect to  $M$  exceeds the elasticity with respect to  $i$  for the observed ranges. A discrete approximation to the elasticities can be calculated with the use of an annuity table. The interest elasticity increases in absolute value with both maturity and interest rates and the maturity elasticity does the opposite. Thus, the comparison of a maturity elasticity of .79 and an interest elasticity of -.21 at a maturity of 36 months and a loan rate of 16 percent does not overstate the case for using the maturity variable. See also Juster and Shay.

<sup>13</sup> Adaptive expectations with a trend correction and the Permanent Income Hypothesis were used to generate the series.

<sup>11</sup> The constructed data together with a discussion of the construction procedures are found in the reprint of this paper which will be distributed by the National Bureau of Economic Research.

in these equations we adjust the lagged dependent variable for such supply influences.<sup>14</sup>

Alternative specifications of the transitory function,  $T$ , are also tested. Unemployed man-hours,  $U$ , as a general measure of cyclical conditions, is preferred.<sup>15</sup> An alternative specification is transitory income proper ( $Y^T$ ), defined as the difference between current and permanent income. However, this variable appears to have only a very gradual impact on investment, which makes it difficult to interpret the lag structure of the model.

Tables 1A, 1B, and 1C present a set of basic regression results for both net and gross investment in total durables, automobiles, and nonauto durables for the 1949-67 period; estimates are by Ordinary Least Squares. The fully specified net investment model ( $AE$ , equation (7) above), utilizes the unemployed man-hours variable as the transitory function. The sign and magnitude tests on the transitory and lagged transitory coefficients are satisfactory.<sup>16</sup> The transitory income variable proper (current less permanent income) did not satisfy the tests; the results indicate a lagged rather than immediate influence on stock change. Rather than complicate the lag structure of the model, this variable is used in the simplified function (equation (8)) described as  $AE-2$ .

The  $AE-1$  equation uses current disposable income as an explanatory variable in a second-order model, while the last two rows  $A$  and  $AT$  of the top panel provide

<sup>14</sup> A discussion of the adjustment procedure is found in an Appendix to the *NBER* reprint of this paper.

<sup>15</sup> Unemployed man-hours are defined as the number unemployed times the average number of hours worked plus hours lost due to involuntary part-time work, divided by the total man-hours of the potential labor force.

<sup>16</sup> An alternative hypothesis—that the first difference in unemployed man-hours is the correct explanatory variable in a model without an explicit transitory component—may be equally plausible.

estimates of a first-order (partial adjustment) model with, respectively, current income and a permanent-transitory distribution of current income. In each of the tables 1A, 1B, and 1C, the five models are shown with both net and gross investment as the dependent variable.

The calculated  $t$ -ratios for the regression coefficients are well above acceptable levels in virtually every instance, and the lag structure in both first- and second-order models are stable. The lagged stock coefficients in the gross investment equations are at times insignificant but there is no a priori reason why these coefficients could not be zero—the adjustment coefficients and the depreciation rate could be of approximately equal size. For the first-order models, the Durbin-Watson statistic suggests that there is positive serial correlation in the residuals.<sup>17</sup>

A closer look at some of the coefficients shows that the maturity variable is not consistently significant in the durables equations. For automobiles, which should be most sensitive to credit changes, the variable is always at least twice its standard error. The coefficients of the price variables exhibit some instability, especially when the unemployment variable is included, probably because of common trends in both variables.

The transitory income coefficient is always highly significant whereas the permanent income coefficient is not, especially for automobiles. The magnitudes of the transitory coefficient in the durables equations are twice that of permanent; for automobiles the ratio is higher and for

<sup>17</sup> This result is common in a quarterly model without a lagged dependent variable. For the second-order models, although the Durbin-Watson is biased towards 2, the results do not preclude the possibility of positive serial dependence. The model was reestimated with the additional assumption that the residuals follows a pattern of first-order serial correlation. The results, which are basically the same as those shown above, are examined in an Appendix to the *NBER* reprint of this paper.

TABLE 1A—ESTIMATES OF REGRESSION COEFFICIENTS OF MODELS OF NET AND GROSS INVESTMENT IN CONSUMER DURABLES<sup>a</sup>  
U.S. Quarterly Data, 1949-67

Model	Intercept	Y	Y*	P <sub>D</sub> /Q	M <sub>D</sub>	S <sub>D-1</sub>	N <sub>D-1</sub>	U	U <sub>-1</sub>	Y <sup>t</sup>	SD	KD <sub>D</sub>	$\bar{R}^2$
<i>Net Investment, N<sub>D</sub></i>													
AET	1104.0		.0485 (2.1)	-897.0 (-3.8)	.7209 (2.8)	-.1587 (-4.7)	.6416 (9.0)	-34.94 (-6.5)	26.16 (5.0)		37.36 (3.9)	93.05 (6.2)	.934
AE-1	214.8	.1093 (4.6)		-429.5 (-1.7)	.5258 (2.5)	-.1744 (-6.0)	.6289 (9.1)				45.19 (4.1)	103.9 (6.4)	.911
AE-2	581.0		.0836 (3.5)	-577.8 (-2.3)	.3265 (1.6)	-.1466 (-5.0)	.5574 (8.4)			.2675 (5.6)	50.10 (4.7)	111.0 (7.0)	.919
A	969.8	.2311 (8.0)		-1105.0 (-3.1)	.3988 (1.3)	-.3477 (-10.5)					39.60 (2.4)	120.7 (5.0)	.804
AT	1290.0		.1729 (5.0)	-1230.0 (-3.7)	.1376 (.5)	-.2783 (-7.9)				.4329 (6.9)	40.22 (2.7)	128.2 (5.7)	.834
<i>Gross Investment, G<sub>D</sub></i>													
AET	804.1		.1193 (5.2)	-944.2 (-4.1)	.7174 (2.8)	-.0113 (-.3)	.6250 (9.4)	-34.52 (-6.5)	28.43 (5.5)		39.67 (4.1)	96.23 (6.5)	.985
AE-1	187.5	.1718 (7.9)		-584.4 (-2.5)	.5813 (3.0)	-.0294 (-1.1)	.5613 (8.8)				47.13 (4.7)	112.6 (7.5)	.983
AE-2	349.0		.1520 (6.7)	-663.7 (-2.9)	.4791 (2.5)	-.0090 (-.3)	.5367 (8.5)			.2617 (5.8)	41.17 (4.8)	116.4 (7.9)	.984
A	778.9	.2805 (10.8)		-1187.5 (-3.6)	.4679 (1.6)	-.1814 (-6.1)					42.13 (2.9)	127.7 (5.8)	.963
AT	1003.0		.2399 (8.2)	-1275.0 (-4.0)	.3074 (1.1)	-.1367 (-4.1)				.4211 (7.1)	42.57 (3.0)	132.7 (6.3)	.966

<sup>a</sup> The independent variables in this regression are as follows:

Y = Disposable income, 1958 dollars per household

Y\* = Permanent income, 1958 dollars per household

P<sub>D</sub>/Q = Relative price of consumer durables

M<sub>D</sub> = Maturity on all consumer durable contracts

S<sub>D-1</sub> = Beginning of period stock of consumer durables, 1958 dollars per household

N<sub>D-1</sub> = Net investment in consumer durables, 1958 dollars per household

U = Unemployed man-hours

Y<sup>T</sup> = Transitory income, 1958 dollars per household

SD = Automobile strike dummy

KD<sub>D</sub> = Korean War dummy for consumer durables

G<sub>D</sub> = Gross investment in consumer durables, 1958 dollars per household

other durables it is about one. Thus there appears to be a strong transitory influence on automobile investment, while nonauto durables are less subject to transitory effects. The permanent income coefficients are always higher in the purchase equations than in the corresponding net investment equations. Transitory income on the other hand seems to effect only net investment and not replacement demand, as the coefficients are unchanged in net and gross investment equations.

The long-run permanent income elasticities implied by the model are all about unity, indicating that the household sector

aims for a constant ratio of durables stocks to income.<sup>18</sup> The equilibrium price elasticities all exceed unity, suggesting that the relatively large secular growth in durable stocks over the last two decades has been largely due to their relative cheapening. Other durables appear to be more sensitive than automobiles to both price and income changes.

The short-run expenditure elasticities implied by the unemployed man-hours variable are quite large, particularly for automobiles. The response is most easily

<sup>18</sup> A full discussion of the statistical properties of the model is found in the *NBER* reprint of this paper.

TABLE 1B—ESTIMATES OF REGRESSION COEFFICIENTS OF MODELS OF NET AND GROSS INVESTMENT IN AUTOMOBILES<sup>a</sup>  
 U.S. Quarterly Data, 1949-67

Model	Intercept	$Y$	$Y^*$	$P_C/Q$	$M_C$	$SC_{-1}$	$NC_{-1}$	$U$	$U_{-1}$	$Y^T$	$SD$	$KD_C$	$\bar{R}^2$
<i>Net Investment, <math>N_C</math></i>													
<i>AET</i>	283.0		.0091 (.9)	-259.4 (-2.8)	.5796 (3.9)	-.1715 (-4.0)	.6752 (9.9)	-25.78 (-5.9)	19.57 (4.4)		44.60 (5.6)	102.5 (2.6)	.879
<i>AE-1</i>	37.79	.0353 (3.2)		-136.3 (-2.3)	.3284 (2.8)	-.1795 (-4.4)	.7109 (9.3)				50.07 (5.4)	120.9 (2.8)	.829
<i>AE-2</i>	202.7		.0210 (2.0)	-241.4 (-2.6)	.4490 (4.3)	-.1735 (-4.8)	.5436 (8.2)			.1926 (5.2)	56.20 (6.8)	125.1 (3.1)	.865
<i>A</i>	574.5	.0601 (3.8)		-673.7 (-4.9)	.5782 (3.3)	-.3527 (-6.4)					45.17 (3.2)	216.1 (3.4)	.611
<i>AT</i>	637.1		.0265 (1.8)	-622.9 (-5.4)	.5390 (3.7)	-.2503 (-5.1)				.3143 (6.6)	46.17 (3.1)	225.9 (4.2)	.731
<i>Gross Investment, <math>G_C</math></i>													
<i>AET</i>	198.9		.0384 (3.5)	-301.8 (-3.0)	.5097 (3.2)	.0046 (.1)	.6843 (9.3)	-22.85 (-5.0)	18.97 (4.0)		50.68 (6.0)	106.3 (2.5)	.934
<i>AE-1</i>	-1.108	.0594 (5.4)		-197.5 (-1.8)	.3198 (2.7)	-.0047 (-.1)	.6891 (9.0)				55.65 (6.0)	131.3 (3.0)	.921
<i>AE-2</i>	85.55		.0488 (4.4)	-239.2 (-2.3)	.3319 (2.9)	.0167 (.4)	.6077 (7.7)			.1646 (4.1)	55.49 (6.3)	146.2 (3.5)	.928
<i>A</i>	519.1	.0834 (5.3)		-718.4 (-5.3)	.5619 (3.3)	-.1726 (-3.2)					50.91 (3.7)	223.6 (3.5)	.826
<i>AT</i>	571.5		.0556 (3.7)	-677.2 (-5.7)	.5276 (3.5)	-.0870 (-1.7)				.2962 (6.0)	51.74 (4.3)	231.8 (4.2)	.864

<sup>a</sup> The independent variables  $Y$ ,  $Y^*$ ,  $U$ ,  $Y^T$ , and  $SD$  are as defined in Table 1A. The others are as follows:

$P_C/Q$ =Relative price of automobiles

$M_C$ =Maturity on automobile installment contracts

$SC_{-1}$ =Beginning of period stock of automobiles and trailers, 1958 dollars per household

$NC_{-1}$ =Net investment in automobiles and trailers, 1958 dollars per household

$KD_C$ =Korean War dummy for automobiles

$G_C$ =Gross investment in automobiles and trailers, 1958 dollars per household

understood as the effect on expenditures of a one point rise in the unemployment rate: such a rise causes an expenditure decline of 6.86 percent for automobiles, 2.11 percent for other durables and 3.86 per cent for total durables.

When the net investment equations are considered as distributed lags in the stock of durables, the mean lags are fairly short but the confidence intervals are wide. However, we can certainly conclude that the mean lags are under one year and are likely to be around four months. First-order models yield somewhat higher mean lags, which is expected.

## II. Models with Anticipatory Data

This section investigates the potential contribution of consumer anticipations

data to models of durables demand. Survey data on consumer attitudes and buying intentions are available at approximately quarterly intervals from 1953 on. The attitudes data (Index of Consumer Sentiment) are a consistent series with the same analytical content and sampling error over the entire period; there are some missing quarters prior to 1961, for which values are interpolated. The intentions data, in contrast, are a spliced series. The only source of such data from 1953 to 1959 is the Survey Research Center (SRC) series, which has both relatively large sampling error, and, in published form, some change in the treatment of responses. From 1959 through 1966 either the SRC series or a conceptually comparable series with much smaller sampling error (the

TABLE 1C—ESTIMATES OF REGRESSION COEFFICIENTS OF MODELS OF NET AND GROSS INVESTMENT IN OTHER DURABLES<sup>a</sup>  
U.S. Quarterly Data, 1949-67

Model	Intercept	$Y$	$Y^*$	$P_H/Q$	$S_{H-1}$	$N_{H-1}$	$U$	$U_{-1}$	$Y^t$	$KD_H$	$\bar{R}^2$
<i>Net Investment, <math>N_H</math></i>											
<i>AET</i>	428.8		.0898 (4.8)	-393.6 (-2.7)	-.2067 (-5.8)	.4301 (4.5)	-9.444 (-3.6)	6.111 (2.3)		92.39	.934
<i>AE-1</i>	91.0	.0944 (7.2)		-181.8 (-2.0)	-.1762 (-7.6)	.4264 (5.4)				99.49 (10.7)	.943
<i>AE-2</i>	91.89		.0941 (5.5)	-182.0 (-2.0)	-.1758 (-6.1)	.4275 (4.9)			.0948 (5.0)	99.51 (10.5)	.942
<i>A</i>	-131.4	.1588 (24.7)		-103.1 (-1.0)	-.2627 (-12.8)					101.2 (9.3)	.921
<i>AT</i>	-142.3		.1669 (20.4)	-107.6 (-1.1)	-.2773 (-12.4)				.1276 (6.1)	99.6 (9.2)	.923
<i>Gross Investment, <math>G_H</math></i>											
<i>AET</i>	738.5		.1003 (5.7)	-715.6 (-5.0)	-.0690 (-2.0)	.4724 (5.0)	-11.24 (-4.4)	8.070 (3.2)		99.0 (9.9)	.990
<i>AE-1</i>	435.5	.1045 (8.0)		-526.2 (-6.0)	-.0416 (-1.8)	.4621 (5.9)				107.0 (11.8)	.992
<i>AE-2</i>	429.1		.1070 (6.4)	-525.0 (-6.0)	-.0453 (-1.6)	.4536 (5.3)			.1013 (5.4)	106.8 (9.8)	.992
<i>A</i>	171.5	.1750 (26.6)		-427.1 (-4.1)	-.1331 (-6.3)					109.1 (9.8)	.987
<i>AT</i>	156.1		.1863 (22.6)	-433.5 (-4.3)	-.1538 (-6.8)				.1306 (6.2)	106.9 (9.8)	.988

<sup>a</sup> The independent variables  $Y$ ,  $Y^*$ ,  $U$ ,  $Y^t$  are as defined in Table 1A. The others are as follows:

$P_H/Q$  = Relative price of other durables

$S_{H-1}$  = Beginning of period stock of other durables, 1958 dollars per household

$N_{H-1}$  = Net investment in other durables, 1958 dollars per household

$KD_H$  = Korean War dummy for other durables

$G_H$  = Gross investment in other durables, 1958 dollars per household

Census Bureau's Quarterly Survey of Intentions (*QSI*) can be used. After 1966 a conceptually different and presumably improved Census series (Consumer Buying Expectations (*CBE*)) is available.<sup>19</sup> We have constructed a continuous series from these sources, using *SRC* data through 1959 and Census data thereafter. The data series used plus a discussion of their con-

<sup>19</sup> The difference among these series are described in Juster (1969b).

struction are detailed in the *NBER* reprint of this paper.

Two general types of demand models that utilize consumer anticipations data are specified. One model views anticipatory data as either substitutes for or complements to the set,  $Z$ , of variables in the desired stock function of the objective model. That is, anticipatory variables can be viewed as additional determinants of desired stock or as substitutes for income,

relative price, etc. as desired stock determinants. An alternative model views anticipatory variables, plans and attitudes, as a possible substitute, not only for the desired stock variables, but also for all lag and adjustment processes specified by the model. This suggests the specification of a pure anticipatory model as a replacement for the objective model and will be discussed first.

### *Anticipatory Models as Substitutes*

Purchase intentions are presumably a direct measure of the difference between beginning of period stocks and planned end of period stocks, hence they could in principle substitute fully for the planned investment part of the objective model. The role of the attitude variable is less clear. One interpretation suggests that intentions are an imperfect measure of the difference between planned and actual stocks and that attitudes serve to modify or correct that measure.<sup>20</sup>

The pure anticipatory model (a gross investment equation) uses intentions  $p$  and attitudes  $A$ , and is designated as  $P$ . Given the specification of the anticipatory variables, the appropriate dependent variable is gross investment measured in physical units purchased, more precisely, the purchase rate ( $x$ ). The  $P$  model is shown as equation (10).

$$(10) \quad x = a_0 + a_1p + a_2A$$

For the anticipatory model with a transitory component, we add  $U$ , the unemployed man-hours variable; the full anticipatory model (11) is designated  $PT$ .

<sup>20</sup> In previous research it has been found that attitudes and lagged intentions were the best predictor of purchase rates or households classed as nonintenders. Hence both intentions and attitudes made significant contributions to an explanation of aggregate purchase rates; intentions presumably reflected variations in intender purchase rates, while attitudes picked up variations in the purchase rates of nonintenders. See Juster (1969a).

$$(11) \quad x = a_0 + a_1p + a_2A + a_3U$$

Comparison of objective and anticipatory models will be facilitated by including several variations of the former in addition to the partial adjustment-adaptive-expectations-transitory change ( $AET$ , equation (7.1)) model outlined above. As the explanatory power of the  $AET$  model may only reflect the existence of serially correlated residuals in an adjustment model, comparisons with the simpler partial adjustment-transitory change model ( $AT$  equation (9.1)) are also made. Another comparison of interest involves the planned investment part of the objective model, that is, the full model without the transitory change component ( $AE$ , equation (9.1)), against the comparable anticipatory model ( $P$ , equation (10)). Since the anticipations models use subjective purchase plans as one of the major ingredients, this comparison answers the question: How well do subjective purchase plans predict behavior relative to their objective counterpart?

Because the consumer anticipations data cover a shorter span than the objective data, comparisons are not possible over the full 1949-67 period used above. They can be made for two shorter time spans, however. The first, 1953-67, involves the longest period for which we have reasonably consistent measures of both consumer attitudes  $A$  and consumer buying intentions  $p$ .<sup>21</sup> The second period covers 1960-67, and is used because it is the only time span for which entirely consistent and statistically reliable measures of both attitudes and buying intentions are available.<sup>22</sup>

The objective model is reestimated for

<sup>21</sup> The intentions variable refers only to automobiles, although it is also used in the total durables function.

<sup>22</sup> As noted above, the available intentions series have differential sampling reliability before and after 1960, while attitudes are not available for every quarter prior to 1961.

each of the two indicated time spans. The anticipations model uses weighted intentions from current and two past surveys  $p$  and lagged consumer attitudes  $A$  to measure planned gross investment; unemployed man-hours  $U$  are used to measure transitory gross investment. Both models are estimated by ordinary least squares, although this procedure may not be entirely satisfactory for purposes of comparison. The objective model includes income and price variables, and the estimates are therefore subject to simultaneous equations bias; the anticipatory model should be largely free of such bias.

The results in Table 2 are interesting, especially where the comparison between objective and anticipations models is unaffected either by large sampling errors in the anticipations variables or conceptual differences between the dependent and independent variables. Both problems are absent in the first two rows of the automobile demand section, where expendi-

tures on automobiles are the dependent variable and the 1960-67 span (when  $QSI$  or  $CBE$  can be used to measure intentions) is the fit period. The objective ( $AET$ ) model performs well in explaining a series with the amount of erratic quarterly variation typical of automobile sales: it explains 94 percent of the variance (adjusted for degrees of freedom); the  $AE$  model, which does not contain the transitory investment variable, explains almost 91 percent of the variance. But the planned investment part of the anticipations model ( $P$ ), consisting only of buying intentions and lagged attitudes, has a slightly smaller standard error than the comparable ( $AE$ ) objective model, and the full anticipatory model ( $PT$ ) has a smaller standard error than the best ( $AET$ ) objective model and a substantially smaller error than the objective model without the lagged dependent variable ( $AT$ ). Thus the much simpler anticipatory models outperform their counterpart objective

TABLE 2—ANTICIPATORY MODELS AS SUBSTITUTES FOR OBJECTIVE MODELS OF DURABLE GOODS DEMAND

Anticipations Model and Time Period	<i>t</i> -Ratios for Anticipatory Models			Standard Errors				
	$p$	$A$	$U$	Anticipatory Model	$PT$	$P$	Objective Model	
							$AET$	$AT$
								$AE$
Automobile Demand								
$P$ 1960-67	+11.5	+2.9	—	—	18.3	—	—	19.1
$PT$ 1960-67	+3.6	+3.9	-3.4	15.5	—	—	16.5	20.3
$P$ 1953-67	+2.7	+2.6	—	—	49.4	—	—	23.2
$PT$ 1953-67	+2.4	+1.4	-5.9	38.4	—	—	19.3	25.9
Durables Demand								
$P$ 1960-67	+14.5	+0.8	—	—	34.4	—	—	21.2
$PT$ 1960-67	+5.2	+1.9	-4.7	25.7	—	—	19.9	22.1
$P$ 1953-67	+6.1	+2.5	—	—	73.4	—	—	26.0
$PT$ 1953-67	+7.0	+1.2	-6.6	54.0	—	—	21.4	29.4

Note: The strike quarters are excluded from the sample period in order to make the standard errors of the anticipatory and objective models comparable. The standard errors are in constant 1958 dollars per household at annual rates. For the anticipatory models, the dependent variables are the automobile purchase rate and a proxy for the durables purchase rate. The variables are both defined as the respective real per household expenditures divided by the average real car price. The data are shown in an Appendix to the *NBER* reprint of this paper. The standard errors for the anticipatory models are adjusted to the same basis as the objective model, as discussed in text: fn. 23.

models.<sup>23</sup> Both intentions and attitudes contribute significantly to the anticipations models, as does unemployed man-hours.

The anticipations model does not fare quite as well in the longer (1953-67) period. For the automobile data, the planned investment objective model is perceptibly better than the anticipations model (compare *AE* with *P*), and the inclusion of transitory stock change improves both models by about the same extent. For the durables equations, the objective model is superior in both periods.

<sup>23</sup> The standard errors shown in Table 2 for the anticipations models are actually obtained from a two-step procedure. Buying intentions are conceptually designed to explain unit sales rather than deflated per household expenditures, hence we estimated the anticipations model with the dependent variable defined as deflated expenditures divided by deflated unit price. This variable is the equivalent of the population purchase rate. The proportion of explained variance and the standard error estimates in the table are not the ones derived from this equation, but rather statistics estimated by multiplying the predicted values from the equation by the deflated price variable. This procedure insures that the standard error estimates for the anticipations models are comparable with those estimated for the objective model.

The durable goods equations are also estimated in the same way, even though the only deflated unit price variable that can be constructed is for automobiles. Thus it is assumed that movements in the deflated unit price of automobiles are identical to those in the deflated unit prices of some weighted average of all durable goods—an extreme assumption but not necessarily a totally unrealistic one.

The anticipations model is a close substitute in the 1960-67 period, especially when the transitory stock change variable is included in both models. For the longer period, the objective models are markedly superior. However, the significance of these results is unclear: they are obtained using an intentions variable that is subject to large sampling error during the 1953-59 period, and that measures only automobile and not total durables buying intentions. On the whole, given the very high standard implied by the content and empirical fit of the objective models, the much simpler anticipations models provide remarkably powerful competition.

### *Anticipatory Models as Complements*

A different but equally interesting question is whether the anticipations variables improve a fully specified objective model, i.e., constitutes a significant subset of the desired stock function. The answer, from Table 3, is unambiguously yes: both buying intentions and lagged attitudes clearly add to the explanatory power of the fully specified *AET* model in the shorter (1960-67) period, both for automobiles and total durables; for the longer (1953-67) period the joint contribution of the two anticipatory variables is not significant, although intentions would be if considered by itself.

Moreover, a modified version of the

TABLE 3—ANTICIPATORY VARIABLES AS COMPLEMENTS TO DURABLE GOODS DEMAND MODELS

Model and Time Period	Standard Errors			<i>t</i> -Ratios for Anticipatory Variables			
	Objec- tive Model	Objec- tive Model Plus	Objec- tive Model Plus				
	( <i>AET</i> )	<i>p, A</i>	<i>p, WDA</i>	<i>p</i>	<i>A</i>	<i>p</i>	<i>WDA</i>
Automobiles, 1960-67	16.5	13.6	13.7	2.2	2.3	2.9	2.2
Automobiles, 1953-67	19.3	19.4	18.2	0.9	0.6	1.9	2.5
Durables, 1960-67	19.9	16.1	17.6	2.5	2.1	2.8	0.8
Durables, 1953-67	21.4	21.1	19.9	1.3	0.7	2.1	2.5

*Note:* Standard errors are in 1958 \$ per household. Variables are described in the text.

consumer attitude variable gives better marks to the anticipatory data. Elsewhere, Juster and Wachtel show that a filtered version of  $A$  appears to provide a better specification of the role of consumer attitudes in forecasting models.<sup>24</sup> The filtered variable, designated  $WDA$ , uses the weighted change in consumer attitude only when it shows either large or persistent change. The results in Table 3 indicate that the  $WDA$  formulation is generally superior to  $A$ , and that both  $p^*$  and  $WDA$ , with one exception, make a statistically significant contribution to the fully specified objective ( $AET$ ) model.

#### *Joint Objective-Anticipatory Models*

The data in Table 3 suggest that the anticipatory variables made a significant contribution to a fully specified objective model, both in the 1960–67 and 1953–67 periods, and for both automobile and total durables expenditure models. Examination of the regression coefficients in a model which simply adds the anticipatory variables to the objective  $AET$  model suggests that even stronger conclusions may be warranted.

In the shorter (1960–67) period, the only variables in the  $AET$  model which retain a  $t$ -ratio in excess of unity, other than the two anticipations variables, are unemployed man-hours and relative price; this finding holds for both automobile and total durables equations. For the longer (1953–67) period, the results are markedly different, possibly because expected purchases are a linked variable containing a great deal of erratic variability in the earlier (1953–59) part of the period. Here, both lagged stock change and permanent income retain statistically significant coefficients in both the automobiles and durables models, while lagged unemployed man-hours is significant in some of

the models. As was true of estimates for the shorter period, the relative price variable lowers the standard error of the model, although its coefficient is never significant at conventional levels.

Although the relative brevity of the 1960–67 period makes it difficult to draw firm conclusions on the matter, it is plausible to conjecture that the optimum specification for a durable goods demand model might well include only the two anticipatory variables, unemployed man-hours, and relative prices. The other two variables that retain explanatory power in the 1953–67 period, permanent income and lagged stock change, are both clearly known to the household at the beginning of the purchase period. Hence, a precise measure of purchase expectations would in principle be expected to eliminate the statistical influence of these two, since purchase expectations should be capable of taking full account of both expected income and all the expectational and adjustment lags specified by the objective model. On the other hand, unemployed man-hours is an integral part of the anticipatory model itself, since it reflects transitory investment, and relative price might plausibly be included as part of the model as well.

The question is whether relative price movements are foreseen or unforeseen at the start of the purchase period. Since the model involves the demand for a class of items that are infrequently purchased, households considering purchase might well be unaware of any recent change in market prices until they begin an active search for the product. Thus if prices have been changing, households may generally tend to be “surprised” at discovering what prices actually are compared to what they had been expecting.

Table 4 presents some regressions which incorporate only those variables which are the best candidates for inclusion in an optimally specified model that combines

<sup>24</sup> See Juster and Wachtel (1972).

TABLE 4—REGRESSION COEFFICIENTS IN DEMAND MODELS COMBINING ANTICIPATORY AND OBJECTIVE VARIABLES

Equation	Constant	$p$	$WDA$	$Y^*$	$P/Q$	$U$	$S_{-1}$	Standard Error	$R^2$
<b>1953-67 Durables Expenditures Dependent</b>									
1	218.5 <i>6.8</i>	26.39 <i>3.6</i>	7.037 <i>4.2</i>		-1660 <i>6.7</i>	-35.24 <i>12.1</i>	+.0856 <i>4.4</i>	23.85	.9808
2	121.4 <i>2.1</i>	31.47 <i>4.1</i>	7.154 <i>4.4</i>	.1289 <i>2.0</i>	-1132 <i>3.2</i>	-26.79 <i>5.2</i>	-.0401 <i>.6</i>	23.15	.9823
3	323.4 <i>12.9</i>	31.19 <i>3.6</i>	5.674 <i>3.0</i>		-2453 <i>12.4</i>	-32.58 <i>9.8</i>		27.95	.9731
<b>1953-67 Automobile Expenditures Dependent</b>									
1	605.0 <i>6.2</i>	23.02 <i>4.1</i>	6.953 <i>4.7</i>		- 480.4 <i>5.0</i>	-15.82 <i>5.4</i>	+.0892 <i>2.9</i>	19.85	.9242
2	582.2 <i>3.7</i>	22.90 <i>4.0</i>	6.903 <i>4.5</i>	.0046 <i>.2</i>	- 468.4 <i>4.0</i>	-15.78 <i>5.4</i>	+.0742 <i>.9</i>	20.06	.9243
3	568.5 <i>5.5</i>	36.25 <i>10.3</i>	4.906 <i>3.5</i>		- 426.3 <i>4.2</i>	-17.12 <i>5.5</i>		21.30	.9109
<b>1960-67 Durables Expenditures Dependent</b>									
1	174.8 <i>1.6</i>	55.43 <i>4.5</i>	5.844 <i>2.2</i>		-1307 <i>1.7</i>	-44.06 <i>5.4</i>	+.0627 <i>.8</i>	16.67	.9912
2	120.5 <i>.9</i>	52.47 <i>4.0</i>	3.872 <i>1.1</i>	.1571 <i>.8</i>	-1034 <i>1.2</i>	-37.03 <i>3.1</i>	-.1360 <i>.5</i>	16.79	.9916
3	266.8 <i>9.9</i>	49.81 <i>4.8</i>	6.061 <i>2.3</i>		-1936 <i>7.3</i>	-46.61 <i>6.1</i>		16.57	.9911
<b>1960-67 Automobile Expenditures Dependent</b>									
1	897.7 <i>1.6</i>	34.77 <i>4.1</i>	5.465 <i>2.4</i>		-544.4 <i>1.3</i>	-29.17 <i>4.2</i>	-.0980 <i>.8</i>	13.85	.9684
2	105.1 <i>1.8</i>	36.12 <i>4.3</i>	7.077 <i>2.6</i>	-.0783 <i>1.1</i>	- 548.4 <i>1.3</i>	-31.83 <i>4.3</i>	+.2044 <i>.7</i>	13.80	.9700
3	482.7 <i>2.6</i>	35.25 <i>4.2</i>	5.447 <i>2.5</i>		- 258.0 <i>1.4</i>	-27.87 <i>4.2</i>		13.74	.9676

Note: The strike quarters are eliminated from all the regressions;  $t$ -ratios are shown in italics. Variables are described in Table 1A or in text.

anticipatory and objective variables.<sup>25</sup> Three equations are presented in each of the four panels: the first two equations are basically partial-adjustment-transitory change models (like  $AT$ , equation

(9.1)) with the anticipatory variables included in the desired stock function; the third assumes that all adjustment processes are represented by the expected purchase variable, as in the  $PT$  model, equation (11), and an additional objective variable is added to the transitory function. The first and second equations differ only in that permanent income is included as a desired stock determinant in the sec-

<sup>25</sup> The joint models are estimated with gross expenditures as the dependent variable. When anticipatory variables are used the appropriate dependent variable is the purchase rate, hence the models may contain specification bias.

ond equation but not in the first. The third equation includes only the anticipatory variables, with relative price and unemployment as the transitory function.

The results support the view that relative prices warrant inclusion in the fully specified model. The best specification for a combined model seems to consist either of eliminating all the adjustment lags and letting expected purchases carry the burden of the adjustment process, or including both expected income and a partial adjustment process in the model; it is not clear which alternative is better. When beginning of period stock is included but expected income is not, the former usually has a positive coefficient: the estimated adjustment coefficient, obtained by the subtraction of depreciation rates from the coefficient of beginning stock, implies a very slow adjustment process. Inclusion of permanent income lowers the beginning stock coefficient and therefore speeds up the adjustment to a more plausible pattern. In the automobile equations, elimination of an explicit adjustment process as well as the expected income variable seems to produce more sensible results than retaining both, while the reverse appears to be true in the durables equations. Needless to say, these conclusions are highly tentative, and are in need of much more exploration.

### III. Summary

On the whole, the evidence suggests that, during periods when both purchase expectations and consumer sentiment can be measured with reasonable precision, the anticipatory model is virtually a perfect substitute for a fully specified objective model, and that as good results can be achieved with a simple two-variable anticipations model as with a much more complex model with a fully specified lag structure. In effect, survey measurements of purchase expectations combined with

systematic changes in consumer sentiment seem able to replace the influence of income and all the adjustment lags in a complex objective model, although it does not appear that the anticipatory variables reflect the influence on purchases of movements in relative prices of durables—possibly because these are largely unforeseen.

The evidence is markedly less convincing during periods when purchase expectations are measured with relatively large sampling errors. Here a significant part of the objective model continues to warrant inclusion in a consumer demand model, and the simple anticipatory model falls considerably short of the fully specified objective model in explanatory power. One clear-cut need for additional research lies in the influence of relative prices on purchase decisions in the context of the model which uses anticipatory variables as the major determinant of desired stock. While most of the evidence seems to suggest that the anticipations variables need to be augmented with a relative price measure, the coefficients of the price variables are erratic and the specification can undoubtedly be improved.

### DATA APPENDIX

For durables, gross investment is Personal Consumption Expenditures on Durables. For automobiles, a quarterly interpolation of expenditures on trailers is added to gross auto product-personal consumption expenditures. This procedure reallocates expenditures on automobile parts to the nonauto durables category. Benchmark stock figures and annual depreciation ratios through 1962 for total durables and the automobile component are from Raymond Goldsmith. Post-1962 depreciation ratios were extrapolated by regression.

The intentions variable ( $p$ ) is constructed from *SRC*, *QSI*, and *CBE* survey sources. For 1953 through 1959, *SRC* data are used; the available raw data are processed, seasonally adjusted, and missing quarters are

interpolated. For 1960 through 1966, *QSI* data are used; *CBE* data are used for 1967. The survey responses are weighted, corrected for the influence of interviewer training sessions and the change in survey type from *QSI* to *CBE*, and then seasonally adjusted. The *SRC* and *QSI-CBE* portions are then linked. The intentions variable weights the current quarter survey value and two lagged surveys by .6, .3, and 1., respectively.

The *SRC* Index of Consumer Sentiment as published in *Business Conditions Digest* is used in lagged form. The filtered attitude variable is:

$$WDA_t = .5D_t(\Delta A_t) + .5D_{t-1}(\Delta A_{t-1})$$

where

$D_t = 1$  if  $\Delta A_{t-i}$  for  $i=0, 1, 2$  are of the same sign

or if  $|\Delta A_t + \Delta A_{t-1}| \geq 7$

or if  $D_{t-2} = 1$  and  $D_{t-1} = 0$  and  $|\Delta A_t| > |\Delta A_{t-1}|$

$D_t = 0$  otherwise

A more complete explanation of the derivation of the data is found in the Data Appendix to the *NBER* reprint of this paper.

# REFERENCES

- G. Chow, *Demand for Automobiles in the United States*, Amsterdam 1957.
- R. W. Goldsmith, *The National Wealth of the United States in the Postwar Period*, Princeton 1962.
- M. J. Hamburger, "Interest Rates and the Demand for Consumer Durables," *Amer. Econ. Rev.*, Dec. 1967, 57, 1131-53.
- S. H. Hymans, "Consumer Durable Spending: Explanation and Prediction," in A. M. Okun and G. L. Perry, eds, *Brookings Papers on Economic Activity 2:1970*, Washington 1970.
- F. T. Juster, (1969a) "Consumer Anticipations and Models of Durable Goods Demand: The Time-Series Cross-Section Paradox Re-examined," in Jacob Mincer, ed., *Economic Forecasts and Expectations*, New York 1969.
- , (1969b) "Consumer Anticipations Surveys: A Summary of U.S. Postwar Experience," paper presented to 9th CIRET Conference, Madrid, Sept. 1969.
- and R. P. Shay, *Consumer Sensitivity to Finance Rates: An Empirical and Analytical Investigation*, Occas. paper 88, Nat. Bur. Econ. Res., New York 1964.
- and P. Wachtel, "Uncertainty, Expectations and Durable Goods Demand Models," in B. Strumpel et al., eds., *Human Behavior in Economic Affairs: Essays in Honor of George Katona*, Amsterdam forthcoming 1972.
- E. S. Maynes, "Consumer Attitudes and Buying Intentions. Retrospect and Prospect," mimeo Sept. 1966.
- E. Mueller, "Effects of Consumer Attitudes on Purchases," *Amer. Econ. Rev.*, Dec. 1957, 47, 946-65.
- R. Stone and D. A. Rowe, "The Market Demand for Durable Goods," *Econometrica*, July 1957, 25, 423-43.
- D. B. Suits, "The Demand for New Automobiles in the United States, 1929-1956," *Rev. Econ. Statist.*, Aug. 1958, 40, 273-80.
- R. Waud, "Misspecification in the 'Partial Adjustment' and 'Adaptive Expectations' Models," *Int. Econ. Rev.*, June 1968, 9, 204-17.

# Disasters and Charity: Some Aspects of Cooperative Economic Behavior

By CHRISTOPHER M. DOUTY\*

Recent investigations by economists and other social scientists into events pursuant to natural disasters have revealed an unexpected pattern of behavior. Economic theory suggests that the sudden, largely unanticipated destruction of wealth by an external force—the characteristics which define a disaster—will lead to a higher price level for necessities. It may also be expected that the cloak of the ensuing mass confusion and uncertainty will result in an increase in all forms of antisocial behavior. Yet, empirical research has repeatedly shown that prices rarely rise enough to clear markets, that natural disasters are typically followed by an increase in charity by residents of the disaster zone, and that there is an increase in “community feeling” generally. The fact of postdisaster charity, and of generally heightened concern for the well-being of others, is of greater theoretical interest than its quantitative importance. Apparently, for some time after a disaster, resources are typically used differently and with more generosity toward others.<sup>1</sup>

\* Christopher Douty died unexpectedly on June 17, 1970 at the age of 33, between the time the paper was submitted for publication and its final acceptance. The sad duty of editing the draft fell to M. W. Reder of the City University of New York as his dissertation advisor and friend. Though a few changes have been made, mainly at the suggestion of referees, the paper is substantially as submitted. Reder writes that “Douty’s untimely death robs the profession of a promising scholar and an uncommonly decent human being. At the time of his death, Douty was assistant professor at the University of San Francisco whose research support he gratefully acknowledged.”

<sup>1</sup> However, an intellectually dissatisfied economist may still derive much emotional satisfaction from these unexpected benevolent actions of human beings under trying circumstances.

This fact presents an anomaly for economic theory to explain. It is the task of this paper to offer such an explanation. It is hoped that the theory advanced also serves to shed light on the “social cement” that normally exists within a community.

The behavior pattern to be discussed has been observed not only after natural disasters, but after virtually all disasters of external origin.<sup>2</sup> Systematic empirical studies of the effects of the 1917 munitions ship explosion in Halifax harbor, of the social effects of World War II bombing, of the events following the 1953 tornadoes at Worcester, Massachusetts and at Waco and San Angelo, Texas, and a study of the 1961 disaster caused by Hurricane Carla have all revealed similar postdisaster behavior.<sup>3</sup> More recently, the economic investigations by Dacy and Kunreuther into the 1964 Alaskan earthquake and a study by Douty of the effects of the 1906 earthquake and fire in San Francisco have turned up essentially the same sequence of events.

This sequence, which Jack Hirshleifer (1963) calls the “disaster syndrome,” has been described in terms of sociological theory by James D. Thompson and Robert W. Hawkes. The community is seen in “normal” periods as a multipurpose system that is organized to enable it to utilize its resources for the achievement of many simultaneous objectives. A disaster is seen as an event that destroys not only wealth but also the allocative and “inte-

<sup>2</sup> Internally generated disasters, such as civil disorders, apparently result in a different behavior pattern.

<sup>3</sup> See Douglas Dacy and Howard Kunreuther, Harry Moore (1958, 1964), and S. H. Prince.

grative" (organizational) mechanisms. Initially, a disaster sharply reduces the interaction—exchange and otherwise—among primary units (i.e., families); this is soon followed by a resurgence of interaction among primary units in relief and rescue operations, but with much greater dependence upon nonmarket coordination than normal. Despite plentiful opportunities, there is virtually no looting or other anti-social behavior; instead the community appears as a "super-organization" with allocative decisions made by a centrally controlled bureaucracy often headed by the predisaster civic leaders. Cooperation and generally selfless behavior by the victims and others near the disaster zone is strikingly evident. However, with the beginning of long-run recovery, individuals resume their normal degree of egocenteredness, with the centralized allocative mechanism either breaking down or withering away.

Discussions of the disaster syndrome have suffered from *ad hoc* theorizing and the difficulty of disentangling reports of events from theories about human behavior under the specified conditions. Several writers anxious to improve on this state of affairs and convinced that economic theory ought to be able to say something about disaster phenomena, have produced the theoretical work that is critically summarized in Section I. Section II develops an alternative theory and its implications are examined.

### I. Some Economic Interpretations of Postdisaster Behavior

An obvious peculiarity of observed post-disaster economic behavior is the failure of prices to rise as rapidly as would be suggested by simple supply and demand analysis. The destruction of the stocks of "necessity goods," of which food, clothing, and shelter are prime examples, might normally be expected to lead to sharply

rising prices during the Marshallian market period (which lasts at least until outside aid is received), as competition for the remaining supplies of necessity goods intensifies.<sup>4</sup> However, social scientists who have studied disasters assert that prices rise less than would be predicted on the basis of predisaster economic relationships and the magnitude of the disaster involved.<sup>5</sup> Apparently, a disaster motivates persons within the disaster zone who have retained undestroyed stocks of necessities to increase their charity. Dacy and Kunreuther and Louis De Alessi have offered theoretical explanations for these phenomena which we shall discuss.

The empirical evidence on postdisaster price behavior is sketchy and impressionistic. Breakdown of communication and transportation obviously fragment markets and increase the expected price dispersion for a commodity within a given geographical area. Records of transactions are few and of uncertain representativeness. Consequently, we shall not attempt to argue that after a disaster prices go up "much" or "little." What is clear is that transactions typically occur at prices that would seem far below what the market would bear. Collateral evidence that this is so, is the prevalence of queues for all sorts of goods with no appreciable tendency for supplies to disappear (to black markets) and queues to rapidly shorten.

In analyzing the behavior of suppliers (donors) in disaster zones, it is useful to divide them into the following groups: 1) households within the disaster zone with accessible undestroyed stocks of necessity

<sup>4</sup> This statement implicitly assumes a high survival rate of the population relative to that of nonhuman wealth and that the victims retain some means of making their offers of exchange for goods and services effective. The former assumption is empirically correct; the latter often is not.

<sup>5</sup> This should not be interpreted as meaning that there are never any price increases or that "extortionate" prices are never observed. See Section II below.

goods; 2) large business firms operating wholly or substantially within the disaster zone; 3) private economic units located in "support" zones; 4) the central and state (or provincial) governments having sovereignty over the stricken area. With the probable exception of category 1) above, the donor is unacquainted with the recipients of his charity; therefore most charity is given to a generalized group called "victims." The analysis of this section applies only to donor categories 1) and 2); in Section II all four categories are considered.

The observed failure of postdisaster prices to rise as sharply as might be expected, indicates to Dacy and Kunreuther, pp. 63-70, that there has been a structural shift in the utility functions of the stricken population. Their analysis assumes, conventionally, that 1) individuals maximize their utility; 2) that the utility enjoyed by each individual depends only on his own consumption of goods and services; 3) business firms seek maximum profits in order to maximize the consumption possibilities of their owners, and 4) philanthropic behavior is inconsistent with profit maximization. These assumptions, commonplace in economic textbooks, imply that a disaster-caused leftward shift of market period supply curves coupled with unchanged demand curves, should result in higher short-period equilibrium prices for necessity goods.<sup>6</sup> Though the extent of these price increases may be mitigated by expectation of outside aid, the failure of prices to rise at all following the disasters considered by Dacy and Kunreuther is seen (by them) as due to "emergent altru-

ism" among those disaster zone residents whose circumstances permit them to offer charity.<sup>7</sup> Thus their explanation of upward price stickiness is that tastes have changed to include more altruism. They allege that at some later date the taste for altruism disappears, thereby "explaining" the fact that the observed increase in community feeling is only temporary.

De Alessi (1967) points out that the explanation offered by Dacy and Kunreuther cannot be empirically refuted because no one has yet learned how to directly observe shifts in utility functions. However, he notes that if "interdependence of utility functions" is admitted, then it is possible to retain the customary assumption that utility functions are unchanged, and yet to develop hypotheses that are, in principle, empirically testable. Interdependence is assumed to imply that individuals feel compassion for those who are less well off materially than themselves. This compassion manifests itself normally in a steady flow of charity from relatively wealthy to less wealthy individuals, with the utility of the donors being maximized when the marginal dollar used for the "purchase" of charity yields an increment of utility equal to that yielded by the marginal dollar spent on any other commodity.

A disaster changes the relative asset positions of the donor and the recipient. Assuming that the predisaster recipient is so unfortunate as to be a disaster victim, the relative deterioration of his position will increase the marginal utility to the donor of a given amount of charity, subject only to the condition that the donor's in-

<sup>6</sup> The decrease in community wealth occasioned by a disaster could also lead to a leftward shift of demand curves, assuming that wealth (or income) elasticities of demand are positive. However, Dacy and Kunreuther believe, pp. 65-66, that in the wake of disaster, purely egoistic concerns would lead to an increase in hoarding, causing the demand curves to shift to the right thereby accentuating upward pressure upon prices.

<sup>7</sup> Chapter 5 of their book contains a fairly extensive presentation of relevant data on the 1964 Alaskan experience and briefer descriptions of observed price behavior following other disasters. However, evidence presented by Douty, ch. 4, on San Francisco's 1906 earthquake and fire indicates that completely stable post-disaster prices are not universal. Although some of the data are unreliable, it is unquestionable that room rents rose substantially.

difference curves be convex. Therefore the donor will increase his flow of charity until he is again maximizing his utility.

The alleged interdependence of utility functions also provides a basis for an explanation of the postdisaster charity given by business firms (see De Alessi (1968)). If the utility functions of individuals who are managers contain nonpecuniary elements, business conduct inconsistent with the presumed goal of the maximization of the present value of the owners' equity will be generated. The tendency for such conduct to exist will be stronger if the owners and managers are different sets of people, than if there is owner management.<sup>8</sup>

The fact that many firms give some charity during nondisaster periods is taken by De Alessi to be evidence that firms may increase their philanthropic activity in postdisaster situations. Furthermore, if the firm counts Good Will among its assets, managerial utility maximization through postdisaster donations of some of the firms' wealth may be consistent with the wealth maximizing interest of the stockholders. The existence of Good Will implies that the market in which the firm sells its products is imperfect. Charity may be offered to help the firm maintain its Good Will. Therefore De Alessi concludes, "...other things being the same, the

smaller the degree of competition, the more post-disaster charity a firm will give" (1968, p. 531).

This argument has considerable appeal because of its simplicity, its ingenious use of conventional economic theory and because it provides testable implications. However, it also has some disturbing implications and leaves unanswered a number of important questions. De Alessi's analysis says little about the characteristics of the donors and the recipients. Perhaps it can be reasonably assumed that all "needy" victims receive aid, from a "spontaneously generated" postdisaster relief organization. Such organizations do in fact often arise, but what causes their emergence? Is it possible to predict who will lead these organizations? Is postdisaster charity given only by individuals who have been benefactors of the relatively poor prior to the disaster?

Let us see what can be said on these matters. In regard to business firms, is the degree of imperfection of competition a relevant factor in determining their post-disaster behavior? Can anything be said about external aid, including that of governments?

## II. A Theory of Postdisaster Cooperation

The proposition that an individual makes those decisions that best serve his own interest, narrowly conceived, has been found to be highly useful in economic analysis. Therefore, it is incumbent upon us to develop a theory of postdisaster charity that shows that "altruistic" behavior may under some circumstances be consistent with enlightened self-interest.

At the risk of belaboring the obvious, we note that interdependence of economic units constitutes much of the subject matter of economic analysis, with market interdependence receiving the greatest emphasis. Theorizing about market behavior has most often proceeded as though

<sup>8</sup> This statement is true if the income of managers is only loosely correlated with their success at maximizing profits. Nonemployee stockholders have the option of selling their stock in firms which from their point of view are managed poorly, a fact that enlarges the potential scope for "noneconomic" managerial behavior still more. (It is presumed here that existing managements can be dislodged only with difficulty.) The managers of a regulated public utility have especially great scope for discretionary behavior; the guaranteed profit rate that their firms typically enjoy means that noneconomic managerial behavior has relatively little adverse effect on the owners' equity position. Therefore the scope of such behavior is limited only by the ability of the regulatory commission to enforce standards of managerial competence. However, any degree of monopoly in the seller's market permits some noneconomic behavior (see Armen Alchian and Reuben Kessel).

the institutional environment were a constant.<sup>9</sup> The institutional environment includes the network of rules and regulations by which a society lives. Some of these are formalized into law; others are self-enforcing customs. The institutional environment is a sort of collective good that can be regarded as the result of a consensual agreement.<sup>10</sup> The agreement need not be universally accepted in all particulars, but it must be generally adhered to if the environment is to be viable. This agreement provides a set of guidelines that specifies the allowable scope of actions that an individual may take in his own interest. Such guidelines are required if there is to be substantial market interdependence and a high degree of specialization according to comparative advantage. The benefits of a stable institutional environment are shared in some degree by all members of the community. A major task of government is to see that these benefits are in fact widely diffused and are not captured entirely by a small minority of individuals within the community who undertake illegal or unethical actions.<sup>11</sup>

<sup>9</sup> Neoclassical economists generally only enumerated the functions of government which they saw as consistent with a *laissez-faire* economic framework. The choice of an actual institutional framework was seen as something apart from economic analysis. In recent years theoretical work on the nature of social choice and collective goods and the related work by Alchian, Harold Demsetz and others on the theory of property rights has introduced a degree of integration of institutional choice and market behavior into the mainstream of economics. Needless to say, the "Institutionalists" (e.g. John R. Commons), did not take the institutional environment as a constant. The same applies for economic historians and economists working in many applied areas, especially in economic development.

<sup>10</sup> Hirschleifer (1967), whose paper has been very helpful in the formulation of these ideas, calls the consensual agreement an "alliance." This term, in turn, has been borrowed from Mancur Olson, Jr.

<sup>11</sup> One of the characteristics of a viable institutional environment is that it must provide for an "equitable" distribution of the benefits of the consensual agreement. If the distribution of benefits is seen by a substantial number of the members of a community (not necessarily a majority) as inequitable, the agreement is no longer sufficiently consensual and will be much more difficult to maintain.

This discussion is relevant in this context because the rapid change in the physical environment that an unexpected disaster imposes causes the predisaster institutional arrangements to become largely inoperative. The degree of the breakdown of institutional arrangements is in large measure a reflection of the magnitude of the disaster and of the uncertainty that it creates. The existence of uncertainty causes individual behavior to be governed by indefinite expectations concerning the future. Among the major factors that determine these expectations are the memory of the predisaster institutional environment and the individual's perception of his current situation in relation to that of the rest of the community. The former includes the individual's predisaster web of social and economic relationships. The latter, of course, will change as the individual gains a clearer perception of his and the overall postdisaster situation. The analysis that follows emphasizes the effects of the continuing reassessment of the postdisaster situation by all of the affected individuals, showing in particular that the widely noted increase in community feeling is the expected outcome of individualistic rational behavior.

However, the increase in community feeling is usually not the first collective postdisaster reaction. The initial reaction is the fragmentation of the stricken community into very small units that have been called "kinship groups."<sup>12</sup> At this point the victims probably observe that a large part—perhaps most—of the predisaster population seems to have survived, although much of the capital stock has been destroyed.<sup>13</sup> However, the individual ignores the larger community until

<sup>12</sup> On fragmentation, see Thompson and Hawkes and the references therein. On the concept of kinship groups, see the references in Olson, pp. 17–18.

<sup>13</sup> The behavior of victims who become totally disoriented by the disaster is not considered here. Most disaster studies indicate that disoriented victims are the exception, rather than the rule.

the survival and well being of persons with whom he had maintained his closest personal ties has been determined. The determination of the survival status of family members can be regarded as the first step in the reduction of postdisaster uncertainty, without which rational decision making is impossible. During this interim period of community fragmentation, market transactions, if any are consummated, are often at elevated prices.<sup>14</sup> However, when the survival status of close relatives is determined, usually through a reunion with them, the period of fragmentation ends. The problem as seen by the individual then becomes the survival of the larger community and the restoration of a viable institutional framework.

Assuming that a supply of "necessity goods" survives the disaster in usable form, these must be distributed in such a manner as to allow the victims to subsist without being forced to evacuate the disaster zone. The charging of elevated prices, or even positive prices, for these goods is inconsistent with the maintenance of the population unless these prices are within the means of all of the victims.<sup>15</sup> A means of nonprice rationing is probably required, since many of the victims are likely to be without accessible liquid assets.<sup>16</sup> The

desire to see a continuance of the community leads to the development of such a rationing system and causes social pressure to be brought to bear on possessors of "necessity goods" who are attempting to "extort" high prices from the victims. These phenomena deserve further attention.

Consider the social pressures that a small grocer with an undestroyed inventory stock faces. He has tomorrow as well as today to think about. His regular customers will be aware of his elevation of prices. The grocer will know that if other grocers are not raising prices, his shortsighted behavior could cost him his clientele—if not immediately, then in the near future. This implies that the desire to maintain continuing economic, as well as personal, relationships will moderate any tendencies to attempt to extract all of the consumers' surplus from buyers. Furthermore, if many of the victims are unable to buy food and if none is donated to them, the grocer could soon find that he is unable to prevent the removal of his merchandise without the formality of payment. Hunger pangs are not to be denied. Therefore, if the grocer perceives that he will receive little revenue even if he attempts to charge sharply higher prices, he may then decide that the rational thing to do is to offer his merchandise free to the victims, on a first come, first served basis. No current revenue is realized by this act of charity, but it may be that there is no long-run sacrifice as the grocer's charity may gain him Good Will in the future. Thus it may be possible to rationalize behavior that appears altruistic as reflecting a long-run

<sup>14</sup> There were many dramatic reports of prices being increased to anywhere from ten to fifty times their pre-disaster levels following the 1906 San Francisco earthquake and fire. See Douty, ch. 4, for references. Recent disasters have produced few such reports, however. The generally smaller magnitude of recent disasters may be responsible.

<sup>15</sup> In an article on nonprice rationing, Tibor Scitovsky defined a shortage as existing when the rise of the price of a commodity would cause it not to be distributed on an "egalitarian basis," or if the charge of any commercially viable price would cause a great inequality of distribution of that commodity. Obviously, this definition can be reasonably applied only to necessities, such as stable foods and shelter. However, these are the types of commodities for which sharp advances in prices would be expected following a disaster if economic individualism was carried to an extreme. A disaster creates shortages in Scitovsky's sense of the term; since these shortages exist, there is a case for nonprice rationing.

<sup>16</sup> It is conceivable that individuals in the disaster

zone who possess accessible liquid assets could lend money to those in need of it. However, in this context, such loans are unlikely to be forthcoming even at very high interest rates because the overwhelming environmental uncertainty would create doubts as to the borrower's earning capacity and hence his ability to repay the loan. Finally there is the possibility that the banks will be closed, causing most of the liquid assets of *all* persons within the disaster zone to be inaccessible.

view of self-interest. A similar rationale may apply to many other acts of kindness commonly observed after a disaster; for example, families taking in homeless friends and relatives (sharing housing) without compensation; donating or lending clothing without reward.

Reconciling this type of behavior with the hypothesis of individual utility maximization—which is maintained—presents an interesting theoretical challenge with several different, though not mutually exclusive answers. First, families, both primary and extended, and networks of friends are linked together in an unwritten mutual insurance plan against economic adversity. The implicit obligation to help others in need is especially strong when the utility loss is great and is manifestly unrelated to any failure on the part of the victim. Natural disasters clearly qualify as situations where “insurance benefits” may be claimed. The obligation to render assistance, upon those able to do so, is enforced *de facto* by the (implied) threat of thereafter being denied insurance protection. As implicit contracts of this sort cannot be enforced through the courts, they tend to arise among people who can exert social pressure upon one another to honor commitments; for example, such implicit contracts are likely to arise among members of an ethnic, racial or religious group. Much altruism is this sort of honoring obligations to members of one’s “mutual insurance club.”<sup>17</sup>

But it is also necessary to account for charity toward total strangers, which is frequently observed in the aftermath of a disaster. Part of the explanation lies in the breakdown of communications and transportation so that the normal sources of aid

from mutual insurers are temporarily cut off; also a clustering of members of an insurance network (not uncommon) may put great strain upon or overwhelm those located elsewhere; this increases the importance of aid from “outsiders.” That is, a natural disaster tends to put pressure upon all members of a community to act as though there were an informal reinsurance society to backstop the informal (primary) insurance networks.

To specify the channels through which this pressure is exerted requires identifying a second spring of altruism; community leadership. Community leaders all have an important stake in maintaining a sense of community values. A sense of community values is manifested by participating in the production or purchase of public goods without being overtly covered; i.e., by not attempting to act as a free rider. Religious leaders benefit from a wide diffusion of a sense of community values in that it facilitates money raising and the contribution of free time. Military leaders benefit by getting more willing soldiers; political leaders by making it easier to collect taxes, enforce laws, get public support for costly public goods, etc. Business leaders obviously have an interest in promoting respect for law and order.

Failure to provide for needy members of the community in time of need engenders cynicism toward community values and resistance to demands made in their name. Conversely, manifestation of community values by provision of disaster insurance promotes belief in an unwritten social (insurance) contract respect for which enhances the utility of community leaders. The ability of leaders to reward those who cooperate with them and punish those who do not is obvious.

Yet a third rationale for postdisaster benevolence is the need for immediate community approval. Normally, the sanctity of life and property are protected by

<sup>17</sup> Of course this is not an attempt to describe the psychic states of charity givers in a postdisaster situation. The argument contends only that the donors act “as if” they were optimizing individual utility in the circumstances in which they find themselves.

the courts and the armed force they can muster to enforce their decisions. In the aftermath of a disaster, courts cannot function and soldiers cannot be mobilized. An individual more fortunate or foresighted than others cannot expect safely to hold his property for what he thinks the market will bear, secure in the knowledge that his property is sacrosanct, regardless of the needs of those around him. On the contrary, his property may be seized and his person severely punished as an engrosser. Security of life and property may depend upon close conformity to community norms of behavior, particularly as regards prices charged. A stricken community has little tolerance for entrepreneurial deviance.

In the terse language of the theory of choice, most individuals own one or more (unwritten) insurance policies which "pay off" when their income falls below a certain minimum level for a socially acceptable cause (e.g., a natural disaster). Conversely, they have a contingent obligation to aid others in danger of falling below the implied minimum for an unacceptable reason. The enforcement of this obligation is by social pressure, subordination to which is necessary to avoid unacceptable risks to life, property and reputation (implicitly related to ability to buy and sell on the same terms as others). Nothing further need be said about the characteristics of anyone's utility function, though, "true altruism" may be present in postdisaster situations.

It may be recalled that De Alessi (1968) concluded that any firm in the stricken area may give charity if its managers have a positive taste for charity. Firms in imperfect competition give more charity than atomistically competitive firms in order to protect their Good Will, as indicated above. However, the postulated taste for altruism is not required if the argument is reframed in terms of the degree of involve-

ment of the firm with the stricken community and the firm's size. Our primary concern here is with large firms, as small firms (i.e., small grocers, etc.) have already been considered. However, it is a fact that the largest private benefactors of a stricken community are large firms that had enjoyed extensive operations in the disaster, but which survive the disaster with many of their assets intact.<sup>18</sup> The location of the surviving assets, whether inside or outside the disaster zone, is not relevant. Size is relevant, however, because it connotes identifiability (particularly if the firm deals with the general public rather than with a specialized group of customers), a significant financial capacity and a degree of Good Will in the markets in which it sells. The firm has an incentive to protect its Good Will if possible. This Good Will is threatened if the firm engages in behavior that the disaster victims regard as unfair. Local executives, whether or not these be the firm's top management, sense this intuitively and therefore extend charity to the victims. Furthermore, they might press for charity from branches of the firm located outside the stricken area. Franchised monopolies are the most vulnerable if they fail to be charitable; "profiteering" at the expense of the victims could cause the firm's franchise to be revoked. The executives can be regarded as part of a "privileged group" whose welfare is tied in a very intimate way to that of the community as a whole. As a result we often observe postdisaster charity not only by large firms, but the executives are frequently prominent in leading the relief bureaucracy. Only if a firm had been suffering chronic losses and planned to go out

<sup>18</sup> Two examples will suffice. The Southern Pacific Company was San Francisco's largest benefactor following the 1906 disaster (Douty, ch. 4). In Alaska in 1964, the Safeway supermarket chain did not increase the prices charged for staple food items, and in a few cases lowered some of them, despite the existence of shortages (Dacy and Kunreuther, pp. 113-18).

of business would it have no self-interest in appearing charitable. In this case, the disaster, by causing an "instant amortization" of the firm's local assets only hastens its departure from the stricken region. However, in most cases, large firms will give more charity than small because of their probable greater financial capacity, their greater identifiability and because they possess more Good Will, than smaller firms.<sup>19</sup> But in cases of both large and small firms, business philanthropy can be explained entirely without reference to altruistic motivations.

Firms that do not have a major commercial interest in the survival of the stricken area are not, in general, important post-disaster donors. Again, this proposition could be tested empirically. Also, the *typical* individual located outside the disaster zone cannot be expected to give charity. Yet disasters do frequently result in a flood of uncoordinated contributions of cash and goods in kind from outside individuals. Some of this charity is undoubtedly given by persons who have relatives or good friends living within the disaster zone.<sup>20</sup> But it cannot be assumed that this ac-

counts for all outside private charity. Some postdisaster philanthropy can be regarded as evidence of pure altruism, motivated by a vague feeling of cultural identification with the victims, and perhaps by the knowledge that their gifts will be put to a use that the donors believe is appropriate.<sup>21</sup> However, only a small portion of the persons residing outside the disaster area do give charity and the size of their individual gifts is usually quite small relative to the donor's assets. It is quite possible that individuals who give postdisaster charity give less than they otherwise would have to other eleemosynary institutions during the relevant time period; i.e., a disaster may have the effect of shaping the timing and form of philanthropy by nonresidents, but not its total amount. Regardless of the empirical validity of this last statement (which would be very difficult to test), the existence of outside charity is extremely important in enabling the stricken community to hold together during an emergency period. While individual outside donations are usually very small, in the aggregate they often bulk very large relative to the immediate needs of the victims.

Despite the regularity of the outpouring of private outside postdisaster charity, federal aid has become increasingly important in more recent disasters. Much of this aid is intended to facilitate private reconstruction, but some emergency aid is extended through the Office of Civil Defense and other agencies. Most of the emergency aid is designed to help maintain order, to assist in the administration of private charity, and in population evacuation if it is required. But the fundamental purpose appears to be improvement of the post-disaster environment. These efforts apparently have the approval and support of the federal government's constituency.

<sup>19</sup> It might even be conjectured that large firms will give relatively more charity than small firms (as a percentage of their tangible assets); however, this is far from certain. In the absence of perfect knowledge, "unreasonable" demands could be made on relatively weak firms that happened to be readily identifiable to a large segment of the victims. The determination of the optimal amount of charity (defined here as the equating at the margin the benefit to the firm in the form of a higher level of Good Will than would otherwise be realized in the postdisaster period with the cost of the goods or revenue donated) is not totally within the donor's range of discretion. Under conditions of uncertainty, the optimal amount of charity could vary for firms of equal financial capacity, raising the possibility that small firms could be called on to give relatively more charity than large firms.

<sup>20</sup> Evidence that there are personal ties between outside residents and the victims is the existence of the "convergence" problem, where communications and other links to the outside world become overloaded. This problem has been noted frequently in sociological investigations. See Charles Fritz and J. H. Mathewson.

<sup>21</sup> The bulk of outside charity to victims comes from individuals sharing the same nationality, implying the importance of cultural identification.

There is no effective way that a private individual living outside the disaster zone can organize an order-keeping force; nor is there any incentive for him to try to do so, as he would share the "benefits" of the order-keeping machinery whether or not he bears any of its costs. The federal government has order-keeping machinery at its disposal and can effectively carry out the wishes of its constituents in this regard.

Delayed federal aid which is aimed at facilitating private reconstruction, is offered because the stricken community has suffered a net loss of wealth for which it is not fully compensated by private charity and insurance payments. If the loss of wealth exceeds the magnitude of private transfer payments (including insurance) from external sources, there may be concern that the predisaster institutional environment cannot be restored. Compassionate feelings for the victims may also have been stirred. As with emergency aid, the incentive for a private individual to extend reconstruction aid is very weak because of an absence of a connection between the cost outlays of the donor and the benefits received. Hence there is a rationale for federal aid.<sup>22</sup>

The argument of this paper refers only to disasters from "external" causes. It does not directly apply to disasters such as civil disorders. In principle, our analysis could be developed so as to apply to postriot behavior as well as postearthquake. However this development would require more conceptual apparatus than we have presented here.

<sup>22</sup> Current federal policy toward disasters has been devastatingly criticized by Dacy and Kunreuther, ch. 9-12; also Kunreuther. The critique is restricted to the method of providing reconstruction aid, which is seen as leading to inequitable treatment of the victims. His point is only to show that some federal aid will be offered following any well-publicized disaster, regardless of the institutional framework for providing that aid. Admittedly, the offer of emergency aid is more certain than of reconstruction aid.

### III. A Concluding Note

It perhaps is not too surprising that societies that experience a severe externally derived source of stress should exhibit an intensified degree of community feeling. Mere threats of destruction by an external agent, such as a threat of aggression by a foreign power, often appear to draw societies more closely together. The conditions for increased community or national solidarity as a result of a threat are that the threat be credible to the populace and that there be an effective consensual agreement concerning the preservation of preexisting law and custom. If or when such a threat materializes, consensual agreement operates to generate superficially selfless actions. The interests of the community and the individual's own self interests are both served by individual philanthropy.

### REFERENCES

- A. A. Alchian and R. Kessel, "Competition, Monopoly and the Pursuit of Pecuniary Gain," in *Aspects of Labor Economics*, Universities-Nat. Bur. Econ. Res. Conference series, Princeton 1962.
- D. C. Dacy and H. Kunreuther, *The Economics of Natural Disasters: Implications for Federal Policy*, New York 1969.
- L. De Alessi, "A Utility Analysis of Post-Disaster Cooperation," in *Papers in Non-Market Decision Making*, Fall 1967, 3, 85-90, Virginia Univ.
- , "The Utility of Disasters," *Kyklos*, 1968, 21, 525-32.
- H. Demsetz, "Toward a Theory of Property Rights," *Amer. Econ. Rev. Proc.*, May 1967, 57, 347-59.
- C. M. Douthy, "The Economics of Localized Disasters: An Empirical Analysis of the 1906 Earthquake and Fire in San Francisco," unpublished doctoral dissertation, Stanford Univ. 1969.
- C. Fritz and J. H. Mathewson, *Convergence Behavior*, Washington 1957.
- J. Hirshleifer, *Disaster and Recovery: An*

- Historical Survey*, RAND Corporation RM-3079-PR, Santa Monica 1963.
- , "Disaster Behavior: Altruism or Alliance?" unpublished paper, 1967.
- F. C. Ikle, *The Social Impact of Bomb Destruction*, Norman, Okla. 1958.
- H. Kunreuther, "The Case for Comprehensive Disaster Insurance," *J. Law Econ.*, Apr. 1968, 11, 133-68.
- H. Leibenstein, "Bandwagon, Snob and Veblen Effects in the Theory of Consumer's Demand," *Quart. J. Econ.*, May 1950, 64, 183-207.
- H. E. Moore, *Tornadoes Over Texas*, Austin, Texas 1958.
- , . . . and the Winds Blew, Austin, Texas 1964.
- M. Olson, Jr., *The Logic of Collective Action*, Cambridge, Mass. 1965.
- S. H. Prince, *Catastrophe and Social Change*, New York 1920.
- T. Scitovsky, "The Political Economy of Consumer's Rationing," *Rev. Econ. Statist.*, Aug. 1942, 24, 114-24.
- J. D. Thompson and R. W. Hawkes, "Disaster, Community Organization and Administrative Processes" in G. W. Baker and D. W. Chapman, eds., *Man and Society in Disaster*, New York 1962.
- A. F. C. Wallace, *Tornado in Worcester*, Washington 1954.

# Optimization and Scale Economies in Urban Bus Transportation

By HERBERT MOHRING\*

A cumulative deterioration of urban mass transportation service—fewer riders lead to less frequent service leads to fewer riders lead to . . .—is frequently noted and deplored. A variety of rather exotic labels has been attached to this phenomenon. William Baumol, p. 425, for example, has referred to it as an example of “dynamic externalities.” The appropriate social response to declining mass transit quality can, I think, more easily be seen by recognizing it to be an example of what happens when demand declines for a commodity whose production involves increasing returns to scale. The purposes of this paper are to justify this assertion and to suggest the magnitude of mass transit scale economies and hence the lower bound<sup>1</sup> for an optimal transit subsidy policy.

Transportation differs from the typical commodity of price theory texts in that travelers and shippers play a producing, not just a consuming role. In using common carrier services, they must supply scarce inputs, their own time or that of the goods they ship, that are essential to the

production process. In dealing with many transportation problems, it is both useful and sound analytically to separate these two roles. That is to say, transport costs can be analyzed as if user inputs are purchased in factor markets rather than supplied in kind. Transport demand can be dealt with as if the price of a trip equals whatever fare is charged *plus* the money value the traveler attaches to the time his trip requires.<sup>2</sup>

Accepting these assertions as valid, Section I briefly discusses the interrelationships among short- and long-run cost schedules, the nature of the subsidy required if short-run marginal cost pricing is to be practiced by an increasing returns activity, and the nature of increasing returns in bus operations. Section II uses this development as a base to present cost models for “steady state” and “feeder” bus routes. Also in this section, cost and related data approximating those which prevail in the Twin Cities metropolitan area are utilized to infer the long- and short-run average and marginal costs of bus trips. In turn, these cost data are used to make rough estimates of the user fares and bus company subsidies that would be optimal *if* it were possible to ignore both the effects of bus pricing on highway congestion and the distortions that would be introduced in the process of raising funds for subsidies in a world in which costless ways of levying lump sum taxes do not exist. An Appendix goes into greater detail on the cost data employed.

\* Professor of economics, University of Minnesota and York University and visiting professor, University of Toronto. I am indebted to Lyn Gerber, director of Scheduling, Twin City Lines, and to John Jamieson, director of Transit Development of the Twin Cities Metropolitan Transit Commission for providing data; to Marvin Kraus, Ann Friedlaender, and Edwin Mills for helpful comments; to Myra Wooders for computer programming and checking algebra; and for financial support, to the National Science Foundation and a grant from the departments of Transportation and Housing and Urban Development to the University of Minnesota Center for Urban and Regional Affairs.

<sup>1</sup> Some important second best justifications for mass transit subsidies also exist. These are not dealt with here.

<sup>2</sup> An excellent demonstration of these assertions is contained in Robert Strotz' “First Parable.”

### I. The Nature of Optimal Subsidies and Bus Route Scale Economies

Suppose that producing widgets requires two inputs, labor and capital, and that the long-run marginal and average cost schedules associated with this process are as drawn in Figure 1. Suppose also that the

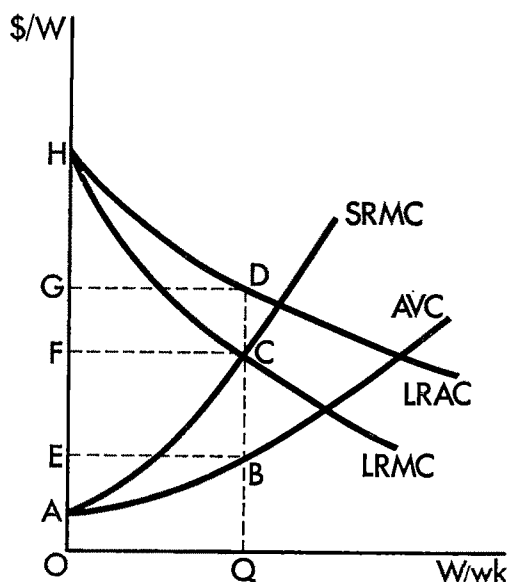


FIGURE 1

widget demand schedule (not drawn to avoid clutter) intersects the long-run marginal cost schedule at C. To minimize the cost of producing  $OQ$  widgets per week would require employing that amount of capital which would generate the short-run marginal and average variable cost schedules shown passing through points C and B, respectively. Setting price equal to long-run (equals short-run) marginal cost would then yield revenues of  $OFCQ$  per week. These revenues would suffice to cover the total costs of labor inputs,  $OACQ$  ( $=OEBQ$ ). In addition, they would yield quasi rents of  $ACF$  ( $=EBCF$ ) on capital inputs. These quasi rents would fall short of the weekly costs of capital inputs. Specifically, since this cost is  $EBDG$  ( $=ACH$ ),

a subsidy of  $FCDG$  ( $=FCH$ ) would be required to cover total costs.

It would make no difference, of course, whether the widget manufacturer used his subsidy check to pay interest on his debts or to cover part of his wage bill. Still, for the purpose at hand, it is important to recognize that the required subsidy equals precisely the amount by which the costs of the fixed inputs he supplies exceed the quasi rents generated in the process of simultaneously minimizing production costs and setting price equal to short-run marginal cost.

Figure 1 can be used without alteration to describe bus operations that are subject to increasing returns. The only difference in interpretation stems from the fact that, with buses, inputs are provided by travelers, not just by the bus company which serves them. Suppose that the number of bus hours of service— $X$ , say—required to minimize the total costs of providing  $OQ$  trips per week has been determined. In this case, “total cost” includes both the cost of providing bus hours and the value consumers attach to the time they must spend traveling. In taking a trip, a representative traveler supplies time valued at  $QB$  dollars. While boarding and alighting, he slows the bus on which he travels. At the very least, the bus door must be left open a few additional seconds to accommodate him. Furthermore, if no one else desires service at his origin or destination, additional stops must be made. Clearly, these delays directly affect those already aboard the bus on which the representative customer travels. In addition, they reduce the number of round trips that can be made with  $X$  bus hours of service, thereby affecting all other bus travelers. The sum of these “own bus” and “system” effects is represented by the line segment  $BC$  in Figure 1.

Since a representative customer provides time inputs valued at  $QB$ , setting the price he pays equal to the short-run margi-

nal cost of his trip would require levying on him a fare—a congestion toll—equal to  $BC$ . If such short-run marginal cost fares were charged,  $EBCF$  would be a quasi rent to bus services, just as in the widget case. In both cases, the required subsidy,  $FCDG$ , is the amount by which this rent falls short of the costs of fixed inputs.

In suggesting the nature of the scale economies which arise in bus operations, it is useful to distinguish between two types of customer supplied inputs—time in transit and waiting time. The latter includes the time a user spends waiting at a bus stop for a bus to come, perhaps waiting at a transfer point for a second bus to arrive, walking from a bus stop to his final destination, and possibly waiting at that destination.<sup>3</sup> If schedules are not published, or if published, not adhered to; if arrivals of buses at major transfer points are not synchronized; and if all trips require specific arrival times and no attempts are made to have buses reach major destinations just before common business opening hours, the average mass transit user could expect a wait of one-half the headway (i.e., the time interval) between successive buses at origin, destination, and transfer points. That is, under these extreme conditions, an expected wait of  $1 + a/2$  times the average bus headway would result if a fraction,  $a$ , of all riders make transfers. Most transit companies do attempt to reduce waiting time in all of the ways suggested. Unfortunately, no evidence is available on the success with which these efforts have met. A wait of half the average headway between buses (a commonly used

number) will be used in all the arithmetic calculations that follow.

Suppose, for sake of illustration, that a bus company initially provides service every 20 minutes on a given route. Suppose, in addition, that the demand for service suddenly doubles and that the company responds (as bus companies commonly do) by doubling the number of buses serving the route. Bus costs *per passenger* would then remain unchanged as would the amount of time a representative passenger spends aboard a bus.<sup>4</sup> However, bus headways and hence waiting time per passenger would both be cut in half *if* the average wait for service is, in fact, proportional to the headway between buses. Thus, the *aggregate* amount of time passengers spend waiting for buses to come would be the same after as before the increase in demand for service. More generally, *if* bus service is provided at a rate proportional to that at which passengers travel, and *if* the average wait for service is proportional to the headway between buses, *then* total waiting time is independent of the number of passengers carried. Under these operating rules, the gap between the average and the marginal costs of a trip equals the value the average passenger places on the time he waits for service. To make marginal cost price viable would, under these circumstances, require providing a subsidy to the bus route equal in value to the stock of consumer-supplied waiting time.

## II. Optimal Tolls and Subsidies in Urban Bus Service

Analysis of Twin City Lines schedules at a sample of points drawn a few years ago from Minneapolis and its immediately adjacent suburbs revealed distributions of elapsed times between buses with means of

<sup>3</sup> Excluding trips to school and home, just over 50 percent of all bus trips in the Twin Cities area had "work" as a destination in 1958, while an additional 25 percent involved social-recreational or personal business purposes (see Minnesota Department of Highways, p. 26). Many trips in these three categories require being at the destination point at a specified time. To the extent that bus schedules do not fit in with such time constraints, a wait at destination is necessary.

<sup>4</sup> If the effect additional bus travel might have on highway congestion and hence on travel time can safely be ignored.

9.3 and 15.6 minutes, respectively, during the peak and off-peak demand periods.<sup>5</sup> Suppose, to pick a number out of the hat, that waiting time is valued at \$1.00 per hour. If the average wait is half the bus headway, the gap between average and marginal costs would then equal about 8 and 15 cents, respectively, during peak and off-peak periods. During the average 1962 weekday, the morning and afternoon peak periods (6:00–9:00 A.M. and 3:00–6:00 P.M.) accounted for 62 percent of total Twin City Lines patronage. Hence a mean gap between average and marginal costs of roughly  $0.62 \times 8¢ + 0.38 \times 15¢ = 10.7¢$  is suggested. The bus company's current basic adult fare is 30 cents. A subsidy per passenger of 10.7 cents would therefore provide a substantial proportionate increase in the bus company's total revenues and hence its capacity to provide improved service.

Actually, the operating rule "make service frequencies proportional to patronage" would not achieve attainable mass transit scale economies. To minimize total costs would require responding to an increase in the demand for service with a less than proportionate increase in service frequency. Both the nature of optimal service characteristics and the rough magnitude of bus route scale economies can be inferred from analysis of some simple bus line cost models. During most of the day on real world bus routes, about the same number of people travel in each direction along a route. During morning peaks, however, substantially more people travel toward than away from the central business district. The reverse is true of afternoon peaks. The numerical calculations described below were carried out for routes

with both balanced and unbalanced flows. However, to make an already very complicated notation as simple as possible, balanced flows are assumed in the descriptions which follow of the "steady state" and "feeder" routes. Given this assumption, if service on one side of a street is optimized, optimization of service on the other side automatically follows.

Consider first a segment of a steady state route. Along each mile of the route:

An average of  $B$  people per hour board and  $B$  exit from buses. Their origins and destinations are uniformly distributed along the route.<sup>6</sup>

$M$  is the length of each person's trip. Hence, at any point along the route segment, an average of  $MB/\chi$  travelers are aboard each bus where

$\chi$  (to be optimized) is the number of buses that traverse the route segment each hour.<sup>7</sup>

$C$  dollars per hour is the cost of providing the services of a bus.

$Y$  (also to be optimized) is the number of uniformly spaced bus stops per mile.

$\gamma$  is the speed at which travelers walk to and from bus stops.

$\beta$  times the headway between buses is the average length of a passenger's wait for service once he reaches a stop.

$V$  dollars is the average value passengers place on an hour spent aboard a bus while

<sup>6</sup> More precisely, origins and destinations are assumed to be uniformly distributed either along the route on which the bus travels or on a continuum of parallel streets which intersect that route.

<sup>7</sup> Actually, the bus company cannot control  $\chi$  but rather only the number of bus hours of service provided to the route, say  $X$ . Because the number of passengers demanding service is a random variable, the number of bus trips that can be provided with  $X$  bus hours is also random. In the analysis, the stochastic nature of the system was, to a considerable degree, ignored. Taking stochastic elements fully into account would have made both analysis and exposition considerably more complex but—hopefully—would not have affected the results appreciably.

<sup>5</sup> In addition to bus company schedules, these data and the numbers which follow were derived from Minnesota Department of Highways, pp. 4, 9–14, 25, 40, and 73–74. See the Appendix for a discussion of how they were developed.

$\alpha V$  is the average value they attach to time spent walking to and from bus stops and waiting for buses to come. Empirical work noted in the Appendix suggests  $\alpha$  to be substantially greater than 1.

$S$  miles per hour is the overall average speed of a bus while

$S^*$  is the speed at which it travels when not engaged in stopping and starting maneuvers.

$\epsilon$  hours are required to board or unload a passenger once the bus has stopped and its doors have been opened.

$\delta$  hours are added to the time required for a bus to traverse the route segment by each stopping and starting maneuver.

The total hourly costs of providing service to an  $M$  mile segment of this route can be broken into four components: bus company operating costs; and the costs to passengers of walking time, waiting time, and time in transit. It takes  $M/S$  hours for the average bus to traverse the route segment. Since  $\chi$  buses per hour do so at  $C$  dollars per bus hour, total bus company costs are  $C\chi M/S$  per hour and the cost per expected passenger served is  $C\chi/BS$ . The distance between stops is  $1/Y$  miles. The maximum walk for any passenger is half this distance. If origins and destinations are uniformly distributed between stops, the average passenger would walk  $1/4Y$  miles both to and from a stop or a total of  $1/2Y$  miles. The cost of such a walk is  $\alpha V/2\gamma Y$  dollars. The average cost per passenger of time spent waiting at a stop is  $\alpha V\beta/\chi$  dollars, while that of time in transit is  $MV/S$  dollars. Summing these four cost components gives the total cost per expected passenger for the steady state route:

$$(1) \quad Z = C\chi/BS + \alpha V/2\gamma Y + \alpha V\beta/\chi + MV/S$$

Suppose, for the moment, that overall bus speed,  $S$ , is independent of  $\chi$ , the rate

at which service is provided. Differentiating equation (1) with respect to  $\chi$ , setting the result equal to zero, and rearranging terms would then yield

$$(2) \quad \chi = [\alpha V\beta SB/C]^{\frac{1}{2}}$$

as the cost minimizing value of  $\chi$ : If speed were independent of level of service, the optimum service frequency would be proportional to the square root of the demand for service.<sup>8</sup>

In fact, if *allowable* stops per mile and passengers per mile-hour are held fixed, a reduction in the number of *actual* stops per mile and hence an increase in realized speed would result from an increase in the number of buses per hour. That is,  $\chi$  and  $S$  are positively related. Thus, the optimal response to a doubling of  $B$  would be to increase  $\chi$  by a factor somewhat in excess of the square root of two.

Determining optimum service characteristics and hence minimum costs, then, requires specifying the relationship between realized speed,  $S$ , on the one hand and, on the other,  $\chi$ ,  $Y$ ,  $S^*$  and the remaining parameters of the system. In each route mile, a total of  $B$  travelers per hour board and  $B$  leave  $\chi$  buses at  $Y$  or fewer stops. Hence, the average number of passengers that board or leave any one bus at any one stop is  $\mu = 2B/\chi Y$ . Suppose that people make travel decisions independently of each other. Then the probability that  $r$  people would board and alight at any one stop is given by the Poisson distribution with parameter  $\mu$ . That is,  $P[r] = e^{-\mu}\mu^r/r!$ . The probability that a given stop will be made, then, is one minus the probability that no one will be at that stop when the bus arrives, i.e.,  $1 - e^{-\mu}$ . The expected number of stops per mile is  $Y$

<sup>8</sup> William Vickrey first propounded this square root principle to me. I presume that he based his assertion on a similar analysis. Actually, Section II of this paper turns out to be merely an elaboration on the top half of page 615 of his 1955 article.

times this fraction. The expected time to travel one mile,  $1/S$ , can therefore be written as the sum of the time actually absorbed in travel,  $1/S^*$ , the time required to board and unload  $2B/\chi$  passengers and the time absorbed by the expected number of starting and stopping maneuvers:

$$(3) \quad 1/S = 1/S^* + 2B\epsilon/\chi + \delta Y[1 - e^{-\mu}]$$

Equations (1) and (3) incorporate the normal operating rule that travelers must walk to more or less widely spaced bus stops. An alternative procedure that is sometimes followed allows a passenger to hail a passing bus at any point along its route at which he happens to encounter it. The prevalent procedure serves to reduce the number of stops a bus makes and hence to reduce both time in transit and the number of bus hours required to provide any specified number of bus trips. At the same time, however, this procedure requires passengers to incur additional walking costs. It seems possible that, on lightly traveled routes, the frequency with which more than one passenger boards or alights at a given stop is so small that the savings in transit time and bus operating costs resulting from limiting the number of allowable stops would not offset the loss in increased walking costs. If so, cost minimization would call for stops to be made on demand. To determine the circumstances under which this possibility might eventuate, the model summarized by equations (1) and (3) was altered to allow for an infinite number of *possible* stops. As  $Y$  approaches infinity, it can easily be shown that equations (1) and (3), respectively, approach

$$(1') \quad Z = C\chi/BS + \alpha V\beta/\chi + MV/S$$

and

$$(3') \quad 1/S = 1/S^* + 2B(\epsilon + \delta)/\chi$$

The second cost model studied deals with a feeder bus route: Along each of the

route's  $M$  miles, an average of  $B$  people per hour board buses. All of them disembark at the route's terminus, downtown. Using the same notation as that for the steady state route (except that  $\mu = B/\chi Y$ ), the cost of serving the passenger who boards at the midpoint of the feeder route can be written:<sup>9</sup>

$$(4) \quad \begin{aligned} Z &= C\chi/BS + \beta\alpha V/\chi \\ &+ \alpha V/4\gamma Y + MV/2S \end{aligned}$$

where:

$$(5) \quad \begin{aligned} M/S &= M/S^* + 2MB\epsilon/\chi \\ &+ \delta[1 + M(1 - e^{-\mu})Y] \end{aligned}$$

With an infinite number of allowable stops, these equations become

$$(4') \quad Z = C\chi/BS + \beta\alpha V/\chi + MV/2S$$

and

$$(5') \quad M/S = M/S^* + MB(2\epsilon + \delta)/\chi + \delta$$

Little would be gained by reproducing the derivatives with respect to  $\chi$  and  $Y$  of any of these relationships. They are quite messy and do not yield explicit relationships for the cost minimizing values of  $\chi$  and  $Y$ . It was therefore necessary to use iterative techniques to find these values.<sup>10</sup> It is possible, however, to find an explicit short-run marginal cost relationship once values of  $\chi$  and  $Y$  have been specified. If  $\chi$

<sup>9</sup> For such a route, it is quite likely that optimum stop spacings and service frequencies would vary with distance from downtown. These possibilities are ignored in what follows.

<sup>10</sup> For reasons suggested by Figure 5, it was impossible to solve simultaneously for the optimizing values of  $\chi$  and  $Y$ . The procedure finally settled upon was that of using Newton's method to determine the optimum service frequency for each of a variety of stop spacings and trip output levels. The long-run marginal costs of providing  $B$  trips an hour was approximated by finding the cost minimizing values of  $\chi$  associated with  $B$  and  $1.05 B$  and then dividing the difference between the two cost levels by  $.05 B$ . Fortunately, costs determined in this fashion typically differed from short-run marginal time costs under optimal conditions by less than half a mill.

bus trips per hour are to be provided,  $X = \chi M/S$  bus hours of service are required. Using this expression to eliminate  $\chi$  in equations (1) and (3), multiplying (1) through by  $MB$  (average passengers per hour in the  $M$  mile route segment of interest), differentiating with respect to  $MB$ , and rearranging terms yields:

$$\begin{aligned} \partial(MBZ)/\partial(MB) \\ (6) &= \alpha V/2\gamma Y + \alpha V\beta M/SX + MV/S \\ &+ 2VM^2B(1 + \alpha\beta/X)A_1/[S(X - 2MBA_1)] \end{aligned}$$

where  $A_1$  equals  $\delta e^{-\mu} + \epsilon$ , the time required to perform a stopping and starting maneuver *times* the probability that the stop at which the marginal expected passenger boards would not otherwise have been made *plus* the time required to board him once the bus has stopped. The term  $2A_1$  is the expected number of hours by which an additional expected passenger would increase the travel time of the  $MB/\chi$  travelers already aboard the bus he takes. Thus,  $2A_1 \cdot V \cdot MB/\chi$  is the cost he imposes on them, the "own bus" effect. The sum of the first three terms on the right of (6) is the travel time cost of a trip—a cost which would be borne by individual travelers. The fourth term, then, is the fare required to equate price and short-run marginal cost. In addition to the "own bus effect," it includes the "system effect" of a trip—the cost it imposes on all other travelers by reducing bus speed and hence the number of bus trips that can be provided by  $X$  bus hours. Equations (4) and (5) yield an *average* gap between short-run marginal and average variable costs for the feeder route of:

$$(7) \quad F = VM^2B(1 + 2\alpha\beta/X)A_2 / [2S(X - MBA_2)]$$

where  $A_2$  equals  $2\epsilon + \delta e^{-\mu}$ , an expression analogous to  $2A_1$ . Equation (7) is the exact marginal cost fare only at the midpoint of the feeder route, where there are an aver-

age of  $MB/2\chi$  passengers aboard the bus. Since no one is aboard a bus when it leaves its outer terminal, the fare there should be  $MBVA_2/2\chi$  *less* than that given by equation (7). On the other hand, short-run marginal cost pricing would require the last person who boards a bus before it reaches its central business district terminal to pay  $MBVA_2/2\chi$  *more* than the amount given by equation (7). That is, optimization of this sort of bus route would require fares to be *inversely* related to length of trip—a perhaps counter-intuitive finding. Figure 4 shows the magnitude of this difference between feeder route short-run marginal cost fares at the route's inner and outer terminals.

In addition to the routes described above on which travel in one direction equals that in the other, solutions were also obtained for routes on which five times as many trips are made in the main direction as in the back haul direction. The data leading to selection of the specific parameter values used are described in the Appendix. These values are:

- $M$  = (Average) trip length: 3 miles
- $\gamma$  = Walking speed: 3 miles/hour
- $\beta$  = Wait for service as fraction of bus headway: 0.5
- $V$  = Value of time in transit: \$1.00/hour
- $\alpha V$  = Value of walking and waiting time: \$3.00/hour
- $S^*$  = Bus speed when not stopping or starting: 20 miles/hour
- $\epsilon$  = Time to board or unload passenger: 1.8 seconds
- $\delta$  = Time to stop and start bus at bus stop: 18 seconds
- $C$  = Cost of a bus hour's services: \$12.75 during morning and afternoon peak; \$5.60 at other times.

The results of these computations are summarized in Figures 2-5 and Tables 1 and 2. Regardless of the specific combina-

TABLE 1—OPTIMUM SERVICE LEVELS AND IMPLIED SCALE-ECONOMY COEFFICIENTS

Period	Peak				Off-Peak	
Flows	Unbalanced		Balanced			
Stops/Mile	8	$\infty$	8	$\infty$	8	$\infty$
Passengers/ Hour/Mile	<u>Steady State Route—Buses/Hour</u>					
150	21.30	45.10	21.20	39.00	41.30	56.60
<i>E</i>	0.57	0.89	0.54	0.86	0.68	0.78
90	15.90	28.60	16.10	25.10	29.10	37.90
<i>E</i>	0.53	0.81	0.54	0.77	0.63	0.77
30	8.90	11.70	8.90	10.80	14.60	16.30
<i>E</i>	0.52	0.67	0.53	0.63	0.57	0.63
9	4.80	5.20	4.70	5.00	7.30	7.60
	<u>Feeder Route—Buses/Hour</u>					
150	25.20	35.40	24.70	31.20	40.90	47.00
<i>E</i>	0.65	0.83	0.66	0.79	0.71	0.79
90	18.00	23.10	17.70	20.80	28.40	31.40
<i>E</i>	0.61	0.74	0.60	0.69	0.64	0.69
30	9.30	10.30	9.10	9.70	14.10	14.70
<i>E</i>	0.55	0.61	0.55	0.59	0.56	0.59
9	4.80	4.90	4.70	4.80	7.20	7.30

TABLE 2—COMPARISON OF OPTIMUM WITH "CURRENT" SERVICE CHARACTERISTICS

	Peak Period				Off-Peak Period	
	Main Haul		Back Haul			
Stops/Mile	16	$\infty$	16	$\infty$	16	$\infty$
	Steady State Route					
Headway	6.0 min.	5.1 min.	6.0 min.	5.1 min.	8.0 min.	7.9 min.
Optimum						
Fare	17.6¢	28.1¢	15.5¢	16.8¢	8.8¢	10.1¢
Subsidy/						
Trip	15.1¢	12.8¢	15.1¢	12.8¢	20.1¢	19.7¢
Price/Trip	63.4¢	70.0¢	51.9¢	47.4¢	50.8¢	48.7¢
	Feeder Route					
Headway	6.2 min.	5.8 min.	6.2 min.	5.8 min.	8.3 min.	8.3 min.
Midpoint						
Fare	13.0¢	16.4¢	9.0¢	9.4¢	5.3¢	5.6¢
Subsidy	15.5¢	14.6¢	15.5¢	14.6¢	20.8¢	20.7¢
Midpoint						
Fare	53.4¢	55.0¢	43.1¢	43.4¢	37.2¢	36.2¢
	"Current" Conditions					
Headway	9.3 min.		9.3 min.		15.6 min.	
Fare	30.0¢		30.0¢		30.0¢	
Subsidy	?		?		?	
Price						
Steady						
State	79.3¢		66.8¢		74.8¢	
Feeder	68.9¢		60.5¢		67.4¢	

tion of parameter values studied, both the steady state and feeder route models reveal considerable scale economies. Operating costs and relative flow rates that currently prevail during peak hours in the Twin Cities mostly underlie Figures 2 and 3. For a two direction average of 150 passengers per mile-hour (250 and 50 in the main and back haul directions, respectively—roughly five times the current peak period average in the area), the long-run marginal costs of main and back haul trips are, respectively, 14 and 19 percent less than overall long-run average costs. Perhaps more important, even for this relatively high output rate, the weighted (by number of trips in each direction) average gap between long-run marginal and average costs amounts to 57 percent of total bus company operating costs. At current travel rates—about 30 and 10 passengers per mile-hour during peak and off-peak periods, respectively—this gap equals 60–61 percent of optimal bus costs.

The differences between the long-run marginal costs of main and back haul trips are surprisingly small as are the differences between the long-run average costs of trips for even and uneven flow conditions. Regarding these latter differences, average costs for even and uneven flows were so close that it was impossible to plot both in Figure 2 except in that range of outputs for which total costs were minimized with an infinite number of allowable stops. This finding is related to the paradoxical results depicted at the bottom of Figure 3: For eight (or fewer) allowable stops per mile, the marginal cost fare for a trip is greater in the back haul than in the main haul direction.

The explanation for these curious results appears to be as follows: The number of passengers assumed to be aboard a back haul bus is a fifth that assumed for a main haul bus. This being the case, an additional back haul traveler has a smaller own bus

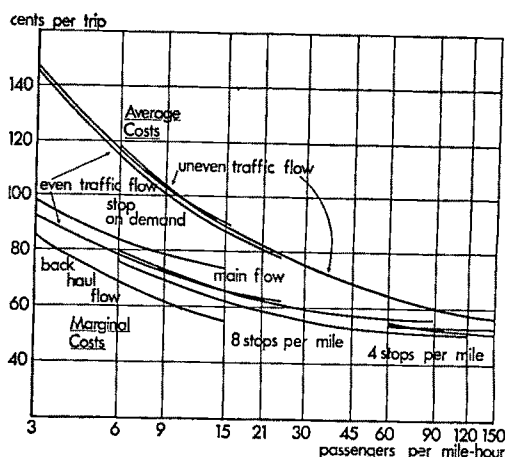


FIGURE 2. STEADY STATE ROUTE:  
MARGINAL AND AVERAGE COST  
(peak period cost condition)

effect on those already aboard the bus he takes than does an additional main haul traveler. But since so few passengers per mile board the average back haul bus, the probability that adding a passenger will require an additional stop to be made is much greater for the back than for the main haul. Under the assumed conditions, it takes 11 times as long (19.8 as opposed to 1.8 seconds) to board a passenger if a special stop must be made for him than it would if the stop would have been made

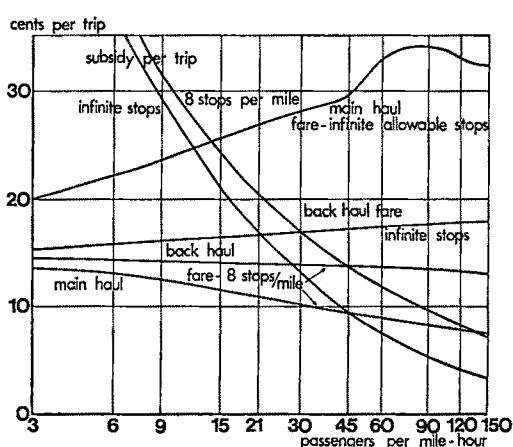


FIGURE 3. OPTIMAL FARES AND SUBSIDIES: STEADY STATE ROUTE, PEAK PERIOD COST CONDITIONS, UNEVEN TRAVEL RATES

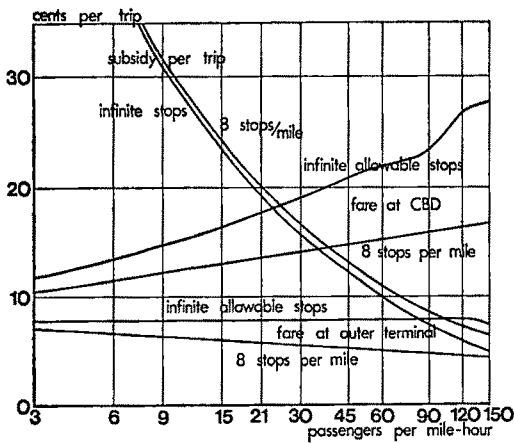


FIGURE 4. OPTIMAL FARES AND SUBSIDIES: FEEDER ROUTE, PEAK PERIOD, EVEN TRAVEL RATES

in the absence of his trip. By reducing operating speed and hence the number of trips that can be made with  $X$  bus hours per hour, additional stops affect all travelers, not just those aboard the bus in question. It would appear that, when the number of stops made per passenger boarded is appreciably less than one, this system effect of an additional back haul trip more than offsets its lower own bus effect.

Going through the arithmetic of a specific example may be worthwhile in this connection. To repeat, during the morning and afternoon peaks, approximately 50 and 10 passengers per mile-hour, respectively, board main and back haul buses on the average Twin Cities route. If stops are spaced an eighth of a mile apart, the cost minimizing service frequency for this output level is 8.88 buses per hour. With this service level, 16.9 and 3.4 passengers, respectively, would be aboard the average main and back haul bus. The own bus effect of an additional main haul trip accounts for 5.8 of the 10.2 cent marginal cost fare; the system effect for the remaining 4.4 cents. The corresponding back haul figures are fare: 13.8 cents, own bus effect: 2.9 cents,

and system effect: 10.9 cents. The substantial difference between the two system effects reflects the fact that the probability that boarding an additional passenger will require an additional stop,  $e^{-\mu}$ , is 0.755 for the back haul but only 0.245 for the main haul.<sup>11</sup>

A point implicit in the foregoing should be made explicit: Stop spacing is a far more important determinant of optimal fares than is the rate at which trips are taken. Thus, as Figure 5 indicates, under peak load cost-even travel rate conditions, marginal cost fares for the steady state route only vary between 2.2 and 3.5 cents over the range 9–150 passengers per mile-hour when one stop per mile is allowed. At the other extreme, fares vary between 19.9 and 26.7 cents over this range of outputs when stops are made on demand.

Although not as dramatic as with fares, stop spacing has a substantial effect on optimal bus headways and hence on the bus operating and travel time components of total costs. These effects are particularly great for high trip output rates. Thus, under peak load cost-even flow conditions, the optimal service level for the steady state route with 150 passengers per mile-hour is 20.4 buses per hour when one stop per mile is allowed, 28.7 for 16 stops per mile, and 39.0 for an infinite number of stops. Bus company operating costs are 10.8, 22.9, and 30.6 cents per passenger under these alternative service rates, and the respective time costs for a three mile trip are 76, 36, and 35 cents.

With few passengers per mile-hour, i.e., for small values of  $B$ , the sum of travel time and bus operating costs is a minimum if buses stop when hailed. For large  $B$  values, between 4–8 stops per mile are op-

<sup>11</sup> If 16 rather than 8 stops per mile are allowed under the conditions dealt with in this paragraph, the paradox would disappear: Marginal cost fares would then be 17.6 and 15.5 cents, respectively, in the main and back haul directions.

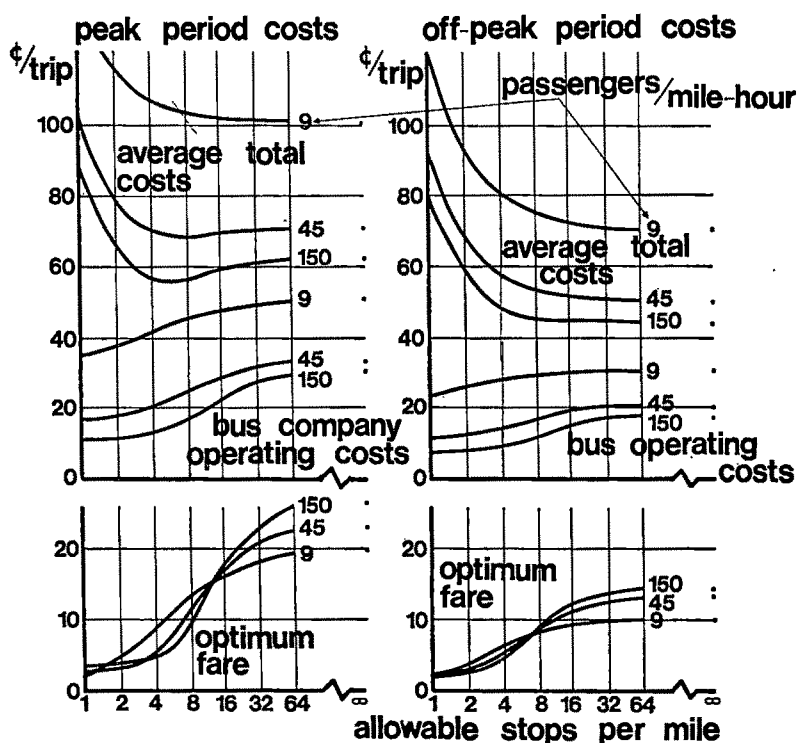


FIGURE 5. AVERAGE TOTAL AND BUS OPERATING COSTS AND MARGINAL COST FARES AT ALTERNATIVE STOP SPACINGS (steady state route, even travel rates)

timal. Finally, for a narrow range of intermediate values, 16, 32, or 64 allowable stops would minimize total costs. Where the dividing lines between "small," "intermediate," and "large" values of  $B$  are drawn depends, of course, on the specific values given system parameters. As the travel time value used decreases, the dividing lines occur at smaller and smaller  $B$  values. On the other hand, decreases in trip lengths and in bus operating costs serve to increase these dividing lines. Thus, for off-peak cost conditions, stopping on demand proved optimal for the steady state route even for the largest value of  $B$  tested, 250 passengers per mile-hour. However, for peak period cost conditions, 4–8 stops per mile provide minimum costs with 45 or more passengers per mile-hour. With  $B$  equal to 30—the present peak

hour average in the Twin Cities—16 stops per mile is optimum, although average total costs for 8, 16, 32, 64, and infinite allowable stops all lie between 74.7 and 75.0 cents.

Table 1 provides data on optimum service frequencies for alternative bus cost, demand, and stop spacing conditions. In addition, denoting alternative  $B$  values by  $B_1$  and  $B_2$  and the associated optimal service frequencies by  $\chi_1$  and  $\chi_2$ , this table gives the value of  $E$  which satisfies:

$$(B_1/B_2)^E = \chi_1/\chi_2$$

That is,  $E$  is the arc elasticity of the optimal service level with respect to travel rate. To repeat, equation (2) indicates that, if the rate at which travelers board a bus has no effect on the speed at which it operates, optimizing service frequencies

would require  $E$  to equal 0.5. If steady state route-peak period costs-8 stops per mile conditions approximate those on a bus system, Table 1 indicates this "square root principle" to be a quite reasonable rule of thumb although an "0.55 principle" would be more nearly accurate. Other combinations of cost and route characteristics yield  $E$  values considerably larger than 0.5. With stop on demand operating rules, a separate stop is made for each passenger regardless of the travel rate. Therefore, for infinite allowable stops, optimum service frequency comes close to being proportional to the demand for service.

The analysis of this paper obviously ignores many problems that would be of relevance in determining desirable fare, subsidy, and operating policies for real world urban mass transit systems. Missing, for example, are consideration of the effects of these policies on traffic congestion, the subsidy implications of the existence of nonoptimal pricing procedures for other forms of urban transportation, and the direct and indirect costs of raising the funds necessary to provide subsidies. Still, it is of interest to compare current mass transit operating and price policies with those this analysis would suggest if all these problems could be ignored. Table 2 constitutes an attempt to do this for conditions currently prevailing in the Twin Cities metropolitan area.

Going from current to optimal operating characteristics would call for the following: a) reducing peak period bus headways about a third; b) reducing off-peak headways by about a half; c) reducing the current 30 cent fare by about 40 percent during the peak period and about 75 percent during the off-peak period. Putting these changes into effect would require increasing the present 600 bus fleet to about 850

and operating about 70 percent of these buses during the off-peak period rather than about 35 percent as is presently the case.<sup>12</sup> If no increase in patronage were to result from these changes, the bus system would generate revenues sufficient to cover about 40 percent of its total costs. However, these changes would reduce trip prices considerably—by about 20 and 30 percent, respectively, in the peak period main and back haul directions and by 35–45 percent during the off-peak period.<sup>13</sup> Patronage would therefore almost certainly increase thereby making further service increases desirable and reducing the system deficit somewhat, at least on a per passenger basis.

To summarize, a subsidy to urban mass transportation systems reflecting the difference between average and marginal costs would probably not eliminate the decline in mass transit usage that has been experienced in virtually all urban areas. After all, currently used mass transit technologies provide services that are inferior goods to most income groups. As real incomes increase, the demand for these services will undoubtedly continue to decline. Still, a welfare maximizing subsidy policy would undoubtedly slow this movement and might even hasten the adoption of new technologies that promise vastly improved service characteristics.

<sup>12</sup> For reasons suggested by the Appendix discussion of the derivation of the \$12.75 and \$5.60 peak and off-peak period costs of a bus hour, an increase in the off-peak fleet utilization rate from 35 to 70 percent would result in both a reduction in the cost of a peak period bus hour and an increase in the cost of an off-peak hour. These changes, in turn, would make optimal somewhat better peak hour service and somewhat poorer off-peak service than levels indicated in Table 2.

<sup>13</sup> The alternative estimates of current trip prices equal the current fare for a trip plus the time costs that would result from the steady state and feeder route models given current travel rates, 9.3 and 15.6 minute headways, and 16 allowable stops per bus mile.

## APPENDIX

*Derivation of Cost and Related Parameters*

Explanations are in order for some of the parameter values used in optimizing bus route service characteristics and in describing current operating characteristics.

A variety of studies (see Michael Beesley, Reuben Gronau, and Thomas Lisco) have concluded that the amounts travelers appear willing to pay to save time aboard mass transit and other vehicles vary with their wage rates. The fraction varies from about 25 percent for low income travelers to about 50 percent for middle and upper income groups. In turn, people appear willing to pay between two-three times as much to save walking and waiting time as to save time aboard conveyances. (See, for example, Lisco, pp. 79-88.) Bus travelers largely come from low income groups. The \$1.00 and \$3.00 an hour used for time in transit and walking and waiting time, respectively, therefore seem reasonable albeit perhaps a bit on the high side.

The amounts \$12.75 and \$5.60 are approximately the marginal costs to the Twin Cities Metropolitan Transit Commission of standard 55 passenger bus hours during peak and off-peak periods. These estimates were determined as follows: A new 55-passenger bus costs approximately \$37,000 and is depreciated over a 12-year period. Applying a 10 percent interest charge to  $\$37,000/2$  and adding depreciation of  $\$37,000/12$  yields an annual capital cost of \$4933.00. Dividing by 1560 peak hours a year—6 peak hours per weekday *times* 5 weekdays per week *times* 52 weeks per year—yields a peak hour capital cost of \$3.16.

Bus driver costs account for about 70 percent of total system costs. Taking into account fringes, overtime, etc., the average cost of a driver hour is currently about \$6.00. Adding a peak bus driver hour involves overtime and other premia that would not be paid for an off-peak driver hour. The Transit Commission's labor contract is extremely complex. It involves, *inter alia*, guaranteed minima of 6-hour days and 40-

hour weeks and time and a half for both hours in excess of 8 per day and beyond 11 hours from starting time. This complexity makes it difficult to determine the exact marginal cost of a peak hour driver. Eight and four dollars for peak and off-peak hours, respectively, seem as reasonable guesses as any. Allocating fuel, tire, maintenance, administrative, and overhead costs on a per bus hour basis adds \$1.58 to the above figures.

Twin City Lines (the predecessor of the Metropolitan Transit Commission) bus trips had mean and median lengths of about 3.5 and 2.3 miles, respectively, in 1962. For the system as a whole during the average weekday, approximately 11 passengers per mile-hour boarded buses between 9:00 A.M. and 3:00 P.M. During morning and afternoon peaks, the average over both directions of flow was about 30. (Derived from Minnesota Department of Highways, pp. 25, 40.) According to bus company personnel, back haul travel during peak hours takes place at the level characteristic of off-peak hours. A two direction average of 30 therefore implies approximately a 50/10 split between main and back haul directions.

In both London and New York, the time required to decelerate, open doors at a stop, and accelerate,  $\delta$ , averages approximately 21 seconds. The value of  $\epsilon$ , is approximately 1.5 seconds for unloading in New York and for both loading and unloading in London. In New York, the time absorbed by fare collection results in a 2.6 second value for  $\epsilon$  in loading.<sup>14</sup>

The estimates of current service characteristics employed in Table 2 and in inferring the magnitude of scale economies under a "service proportional to demand" policy were determined as follows: A sample of

<sup>14</sup> I am indebted to T. M. Coburn of the United Kingdom Road Research Laboratory for these data and also for noting that two-door operation characterizes most American bus services. As a result, total loading and unloading time at a stop is approximately equal to the greater of 2.6 seconds *times* the number of boarding passengers and 1.5 seconds *times* the number of dismounting passengers. This complicating factor is ignored.

points was drawn from the 28 traffic analysis districts covering Minneapolis and its immediately adjacent suburbs that were employed in the 1958 Twin Cities Area Transportation Study. Almost all (95 percent) of the bus trips taken during the survey period had "home" as either origin or destination. The number of observations drawn from each district and the weights attached to each observation were therefore based on the number of trips originating at home that did not have "school" as a destination. School destination trips were eliminated because the majority of them are taken in other than mass transit buses.

Approximately 65 percent of all nonschool trips taken in the survey area had an origin or destination in either the St. Paul or Minneapolis central business district and most bus routes either terminate in or go through one of the Central Business Districts (CBD). Peak and off-peak service frequency estimates for each sample point were therefore based on the number of buses scheduled toward the CBD during the period 6:00–9:00 A.M. and 9:00 A.M.–3:00 P.M., respectively, on the nearest street having bus service. In the few cases where the nearest bus route did not provide service to the Minneapolis CBD, the buses counted were those heading away

from the nearest terminal. See Minnesota Department of Highways, pp. 4, 9–14, 73–74.

#### REFERENCES

- W. Baumol, "Macroeconomics of Unbalanced Growth: The Anatomy of Urban Crisis," *Amer. Econ. Rev.*, June 1967, 57, 415–26.
- M. Beesley, "The Value of Time Spent in Traveling: Some New Evidence," *Economica*, May 1965, 32, 174–85.
- R. Gronau, *The Value of Time in Passenger Transportation: The Demand for Air Travel*, New York 1970.
- T. Lisco, "The Value of Commuters' Travel Time: A Study in Urban Transportation," unpublished doctoral dissertation, Univ. Chicago 1967.
- R. Strotz, "Urban Transportation Parables," in J. Margolis, ed., *The Public Economy of Urban Communities*, Washington 1965, 127–69.
- W. Vickrey, "Some Implications of Marginal Cost Pricing for Public Utilities," *Amer. Econ. Rev. Proc.*, May 1955, 45, 605–20.
- Minnesota Department of Highways. *The Role of Mass Transit: Twin Cities Metropolitan Area*, St. Paul 1963.

# The Economics of Environmental Preservation: A Theoretical and Empirical Analysis

By ANTHONY C. FISHER, JOHN V. KRUTILLA, AND CHARLES J. CICHETTI\*

Concern over the adequacy of nature's endowments has been reflected in economic literature at least from the time of Malthus. For Malthus, the natural environment was essentially a source of increasingly scarce resources to sustain economic activity. Recent theoretical contributions in this framework have sought to develop programs for the optimal intertemporal consumption of fixed and renewable natural resource stocks.<sup>1</sup> Some evidence, on the other hand, suggests that technological progress has so broadened the resource base that the scarcity foreseen by Malthus and assumed, for example, in the stationary utility function postulated by Plourde, has not in fact been realized.<sup>2</sup> Yet, though the statistical evidence is that the direct costs of production from natural

resources have fallen (relatively) over time, it seems likely that some of the environmental costs have risen.

It is desirable to distinguish two kinds of environmental costs. One is pollution, concerning which there is a relatively large and growing literature,<sup>3</sup> which we do not address in this paper. The other is the transformation and loss of whole environments as would result, for example, from clear cutting a redwood forest, or developing a hydroelectric project in the Grand Canyon. Surely there are important economic issues here, yet although there is a vast literature dating back to the 1930's on benefit-cost criteria for water resource projects, economists have said virtually nothing about the environmental opportunity costs of these projects. Where reference is made to the despoliation of natural environments, note is made only in passing to "extra-economic" considerations.<sup>4</sup> Similarly in the texts on land economics no mention is made of the economic issues involved in the allocation of wildlands and scenic resources, nor do the costs of land development include the opportunity returns foregone as a result of destroying natural areas.

More recently Krutilla has argued that private market allocations are likely to preserve less than the socially optimal

\* Fisher's work was done at Brown University and Resources for the Future, Inc. Krutilla and Cicchetti are at Resources for the Future, Inc. This paper represents work done in the Natural Environments Program, Resources for the Future, Inc. Fisher's work was additionally supported partially by NSF Grant GS2530 to the Institute for Mathematical Studies in the Social Sciences, Stanford University. We are indebted to George Borts, John Brown, and Harl Ryder for many perceptive comments and suggestions. We are also grateful to our colleagues at Resources for the Future; to faculty and students of the Natural Resources Institute, Oregon State University 1969, and to Darwin Nelson, Arnold Quint, and Donald Sander of the Federal Power Commission for many constructive suggestions. We wish to acknowledge as well comments on an earlier draft of this paper from Gardner Brown, Ronald Cummings, A. Myrick Freeman III, Richard Judy, Clifford Russell, V. Kerry Smith, and an anonymous reviewer.

<sup>1</sup> See for example, studies by Vernon Smith, Charles Plourde, Oscar Burt and Ronald Cummings.

<sup>2</sup> See the studies by Neal Potter and Francis Christy, and Harold Barnett and Chandler Morse.

<sup>3</sup> For a summary of this literature, see E. J. Mishan.

<sup>4</sup> See for example, *Proposed Practices for Economic Analysis of River Basin Projects*, p. 44, Krutilla and Otto Eckstein, p. 265, Roland McKean, p. 61, and Maynard Hufschmidt, Krutilla, and Julius Margolis, pp. 52-53.

amount of natural environments. Moreover, he concludes that the optimal amount is likely to be increasing over time—a particularly serious problem in view of the irreversibility of many environmental transformations.

In this paper we extend Krutilla's discussion in two ways. First, in Sections I and II we develop a model for the allocation of natural environments between preservation and development. Then, in Section III, we apply the model to a currently debated issue: Should the Hells Canyon of the Snake River, the deepest gorge on the North American continent, be preserved in its current state for wilderness recreation and other activities,<sup>5</sup> or further developed as a hydroelectric facility?

## I

Before proceeding with the discussion of allocation between preservation and development, we observe that a natural area may have not just one, but several uses in each state. For the development alternative, we abstract from this problem by assuming allocation to the highest valued use or combination of uses via the market, or some appropriate mix of market, government intervention and bargaining.<sup>6</sup> Similarly for an area reserved from development, we make the same assumption; i.e., the area is used optimally for recreation.<sup>7</sup> Our objective at this stage, then, is

<sup>5</sup> For a discussion of some of the uses of a preserved natural environment, including some suggestions as to how benefits might be evaluated, see Krutilla.

<sup>6</sup> At least two types of externality, pollution and crowding, are likely to be significant in the commercial exploitation of a natural area, making an efficient allocation in general unattainable in the absence of some form of government intervention or private bargaining to internalize. For a summary discussion of the general externality problem, see Mishan. For an interesting treatment of the crowding problem in particular, see Smith.

<sup>7</sup> The problem is that beyond some point, expanding recreation activity can result in congestion disutility to

to formulate a model for guiding choice between the two broad alternatives of preservation and development.

We begin in this section with a rather general model for the optimum use of natural environments. In succeeding sections a more specific methodology will be developed and used to evaluate the Hells Canyon alternatives.

As a defensible definition of optimum use we propose that use which maximizes the present value of net social returns, or benefits, from an area. In symbols, we wish to maximize

$$(1) \quad \int_0^{\infty} e^{-\rho t} [B^P(P(t), t) + B^D(D(t), t) - I(t)] dt$$

where  $B^P$  and  $B^D$  are expected net social benefits (benefits minus costs) at time  $t$ , from  $P$  units of preserved area, and  $D$  units of developed;  $I$  is the "social overhead" capital investment cost at time  $t$  of transforming from preserved into developed; and  $\rho$  is the social discount rate. Note that the opportunity costs of development, the benefits  $B^P$  from preservation, generally ignored in benefit-cost calculations, here enter explicitly into the expression to be maximized.

There are several constraints, imposed by nature and past development, on the maximization of (1). We recognize, first, that the amount of any given area developed, residually determines the amount preserved. In symbols,

$$(2) \quad P + D = L$$

where  $L$  is the fixed amount of land in the area.<sup>8</sup> Second, current and future choice is

recreationists, or ecological damage, or both. For a detailed discussion, see Fisher and Krutilla.

<sup>8</sup> For sufficient flexibility in application, we think of  $D$  as the number of units affected by the development activity, adjusted perhaps for the character of the activity.

constrained by the results of past choices. In symbols,

$$(3) \quad P(0) = P_0 \quad \text{and} \quad D(0) = D_0,$$

i.e., initial values for preserved and developed portions of the area are given. The dynamic and irreversibility constraints are:

$$(4) \quad D = \sigma I,$$

where  $\sigma$  is a positive constant of transformation with dimensions area/money,<sup>9</sup> and

$$(5) \quad I \geq 0$$

Clearly, were the converse true, i.e., were the transformation reversible, much of the conflict between preservation and development would vanish. It seems to us that it is precisely because the losses of certain natural environments would be losses virtually in perpetuity that they are significant.

Finally, we assume concave benefit functions  $B^P$  and  $B^D$ , so that returns to increasing preservation or development are positive but diminishing; in symbols,

$$(6) \quad B_P^P, B_D^D > 0 \quad \text{and} \quad B_{PP}^P, B_{DD}^D < 0$$

It is conceivable that initial stages of water resource development may be characterized by increasing returns. This will not in general be true of river systems in advanced stages of development, such as the Columbia River system, of which the Hells Canyon reach of the Snake River is a part. Accordingly, while the larger High Mountain Sheep project is more profitable than the smaller Pleasant Valley-Low Mountain Sheep, any increase in scale beyond High Mountain Sheep runs into

severely diminishing returns, as the higher pool reduces the existing developed head upstream. Moreover, though this anticipates the analysis just a bit, what really matters is the behavior of development benefits *net* of opportunity costs. And the marginal opportunity costs of development, the benefits from preservation, are *increasing* as *development* increases.<sup>10</sup>

We now proceed with a control-theoretic solution of this problem in the general case, in which no restrictions are placed on the time paths of the benefit functions.<sup>11</sup> The Hamiltonian is

$$(7) \quad H = e^{-\rho t} [B^P(P, t) + B^D(D, t) - I(t)] + p(t)\sigma I(t)$$

where the first term on the right-hand side,  $e^{-\rho t} [B^P(P, t) + B^D(D, t) - I(t)]$ , is the (discounted) flow of net benefits at time  $t$ , and  $p(t)$  is the (discounted) shadow price (value of future benefits) of development. Setting  $q(t) = \dot{p}(t)\sigma - e^{-\rho t}$ ,  $H$  can be simplified to

$$(8) \quad H = e^{-\rho t} [B^P(P, t) + B^D(D, t)] + q(t)I(t)$$

Note the relationship of  $q$  to  $p$ . If technology or demand relationships are changing, then  $p$  and hence  $q$  will be affected.

Applying the maximum principle of Pontryagin, et al.,  $I$  is chosen to maximize  $H$  subject to the irreversibility restriction (5):

$$(9) \quad \begin{aligned} H \text{ is maximized by } & I = 0 \quad q < 0 \\ & I \geq 0 \quad q = 0 \end{aligned}$$

For  $q > 0$ , investment would have to be infinite over an interval. Quite apart from its impracticality, this possibility can be ruled out because it leads to a contradic-

<sup>9</sup> In specifying the constraint in this fashion we are assuming "constant returns" to increasing investment. This seems at least as plausible, in the general case, as either increasing or decreasing returns, as would be implied by some more complicated functional form for the relationship between investment and development.

<sup>10</sup> This follows from the other half of equation (6), namely that  $B_P^P > 0$  and  $B_{PP}^P < 0$ .

<sup>11</sup> Problems similar in form to (1)-(6) have recently been studied by Arrow and Kurz and by Arrow. In the remainder of this section we draw heavily on their work.

tion. Obviously, past development could not have been optimal; more should have been invested earlier.

Since, from (2)  $P$  and  $D$  are not independent,  $H$  can also be written

$$(10) \quad H = e^{-\rho t} [B^P(L - D, t) + B^D(D, t)] + q(t)I(t)$$

Again applying the maximum principle,

$$(11) \quad P = - \frac{\partial H_{\max}}{\partial D} = - e^{-\rho t} (-B_P^P + B_D^D)$$

Since equation (9) is written in  $q$ , not  $p$ , let us write

$$(12) \quad \begin{aligned} \dot{q} &= \sigma P + \rho e^{-\rho t} \\ &= \sigma e^{-\rho t} (B_P^P - B_D^D) + \rho e^{-\rho t} \\ &= e^{-\rho t} [\rho - \sigma (B_D^D - B_P^P)], \end{aligned}$$

and,

$$(13) \quad \begin{aligned} q(t_1) - q(t_0) &= \int_{t_0}^{t_1} e^{-\rho t} [\rho - \sigma (B_D^D - B_P^P)] dt \end{aligned}$$

From (9), the optimal development path is a sequence of intervals satisfying alternately the conditions  $q(t)=0$  and  $q(t)<0$ . Following Kenneth Arrow, define intervals in which  $q(t)=0$  as free intervals, intervals in which  $q(t)<0$  as blocked (no investment) intervals. In a free interval,  $\dot{q}=0$ , so

$$(14) \quad \rho = \sigma (B_D^D - B_P^P)$$

Assume, however unrealistically, that investment were costlessly reversible, except for prior interest charges. This would be equivalent to renting the area for this period, at a rate equal to the rate of interest. As in the related capital accumulation problem, optimal investment policy would then have the myopic property

$$(15) \quad \frac{\rho}{\sigma} = B_D^D(D^*, t) - B_P^P(P^*, t),$$

or

$$B_D^D(D^*, t) = \frac{\rho}{\sigma} + B_P^P(P^*, t),$$

which may be interpreted to mean that optimal investment policy equates the marginal benefits from development  $B_D^D$  to the sum of direct and marginal opportunity costs  $(\rho/\sigma + B_P^P)$  at any point in time.

Combining (14) and (15), we have

$$(16) \quad D(t) = D^*(t) \text{ on a free interval}$$

Again, following Arrow, define a rising segment of  $D^*(t)$  as a riser. Then, since  $D(t)$  is increasing on a free interval,  $D^*(t)$  is increasing, and a free interval lies within a single riser.

On a blocked interval  $(t_0, t_1)$ ,  $0 < t_0 < t_1 < \infty$ , it follows that  $D(t_0) = D^*(t_0)$  and  $q(t_0) = 0$ , since  $t_0$  is also the end of a free interval. Since  $I = 0$ ,  $D(t)$  is constant, so  $D(t) = D^*(t_0)$ ,  $t_0 \leq t \leq t_1$ . Similarly, since  $t_1$  is the start of a free interval,  $D(t) = D^*(t_1)$ ,  $t_0 \leq t \leq t_1$  and  $q(t_1) = 0$ . Summarizing, on a blocked interval  $(t_0, t_1)$ ,  $0 < t_0 < t_1 < \infty$ ,

$$(17) \quad D^*(t_0) = D^*(t_1),$$

$$(18) \quad \int_{t_0}^{t_1} e^{-\rho t} r[D^*(t_0), t] dt = 0$$

where

$$r(D, t) = \rho - \sigma [B_D^D(D, t) - B_P^P(P, t)],$$

$$(19) \quad \int_{t_0}^t e^{-\rho t} r[D^*(t_0), t] dt < 0, \quad t_0 < t < t_1,$$

and

$$(20) \quad \int_t^{t_1} e^{-\rho t} r[D^*(t_0), t] dt > 0, \quad t_0 < t < t_1$$

Equations (18)–(20) can be given eco-

nomic interpretations. Holding  $D(t) = D^*(t_0)$ , net marginal benefits ( $B_D^D - B_F^D$ ) first exceed (constant) marginal costs, since we do not invest, or push development, to the point ( $D^*(t)$ ) at which they are equal. As short-run optimal development ( $D^*(t)$ ) begins to fall, however, beyond some point there is too much development, i.e.,  $D(t) = D^*(t_0) > D^*(t)$ . From this point, marginal benefits are less than marginal costs. Equation (18) then says that over the full interval ( $t_0, t_1$ ) the sum of (discounted) marginal costs just equals the sum of (discounted) marginal benefits. Equation (19) says that, over an interval starting at  $t_0$  and ending at any time  $t$  short of  $t_1$ , marginal benefits exceed marginal costs. Equation (20) is, of course, not independent of (18) and (19), and says that, over an interval starting at any time  $t$  beyond  $t_0$  and ending at  $t_1$ , marginal benefits are less than marginal costs.

Myopic ( $D^*$ ) and "corrected" ( $D$ ) optimal development paths are shown in Figure 1. Note that at a point such as

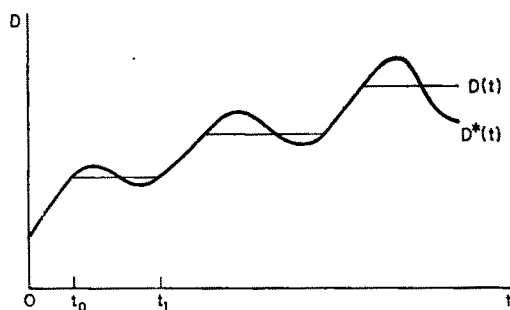


FIGURE 1

$t_0$  at which  $D^*$  is rising, if it will be falling in the relatively near future, then the present value of benefits may be sufficiently low for  $q < 0$ , and investment should cease (equation (9))—until  $t_1$  (equation (17)). We should observe, then, an alternating sequence of rising segments and plateaus in the path of optimal growth over time of the stock of developed land.

The divergence of this corrected path from the myopic is a crucially important result. It says that it will in general be optimal to refrain from development even when indicated by a comparison of current benefits and costs if, in the relatively near future, "undevelopment" or disinvestment, which are impossible, would be indicated.<sup>12</sup>

## II

In the foregoing analysis no restrictions were placed on the patterns of time variation of the benefit functions. But when we come to consider the Hells Canyon project, and quite probably other similar proposals, both theoretical and empirical considerations suggest that benefits from development are likely to be decreasing, whereas benefits from preservation are likely to be increasing. The former, at least, may seem implausible. After all, shouldn't the demands of a growing economy increase the benefits from development of a natural area such as a hydroelectric power site? In this section we first explore this question, and the related one concerning the time pattern of benefits from preservation, then go on to show how the suggested restrictions affect optimal policy.

The traditional measure of the benefits of a hydroelectric power project, at any point in time, is simply the difference in costs between the most economic alternative source and the hydro project. This assumes, of course, that the amount of power provided by the project will be provided in any event, so that gross benefits are equal and the net benefit of the project is the saving in costs.<sup>13</sup> However, over the relatively long life of a hydro

<sup>12</sup> This result was anticipated by Krutilla, who noted "... our problem is akin to the dynamic programming problem which requires a present action (which may violate conventional benefit-cost criteria) to be compatible with the attainment of future states of affairs" (p. 785).

<sup>13</sup> For a fuller discussion of this point, see Peter Steiner.

project, costs of the (best) alternative source of energy will be decreasing as plants embodying new technologies replace the shorter-lived obsolete plants in the alternative system. This means that the benefits from developing the hydro project are correspondingly decreasing over the life of the project. In the traditional benefit-cost analysis this adjustment is not made. Benefits are calculated as of the construction date, implicitly assuming that the technology of alternative sources is fixed over the entire life of the project. For purposes of discussion in this section, a simplified process of technical change and replacement involving some constant rate of decrease of benefits is considered. The implications of a more complicated and realistic process are derived in the Appendix, and applied in our computations in the next section.

Benefits from not developing, on the other hand, appear to be increasing over time. The benefit from a nonpriced service such as wilderness recreation in the Hells Canyon, at any point in time, is the aggregate consumer surplus or area under the aggregate demand curve for the service. Much evidence suggests that demand for wilderness recreation in general, and for the Hells Canyon area in particular, is growing rapidly. This growth is due perhaps to growing population and per capita income, with the extra income used by consumers in part to "purchase" more leisure for themselves. Rising education levels, which seem to be associated with increasing preferences for taking this leisure in a natural environment doubtless also account for the rapidly growing demands.<sup>14</sup> Growth in demand can be broken down into two components: a quantity and a price shift. The effect of

population growth, for example, given unchanging distributions of preferences and income, would be to increase the quantity demanded by the same percentage at any given "price," or willingness-to-pay.

On the other hand, for any fixed quantity, assuming growth of incomes, a set of conditions which will guarantee an increase in price to occur can be summarized as follows: if (a) present services of the environmental resource have no good substitutes among produced goods, (b) income and initial price elasticities of demand for such services are larger than for produced goods in general, and (c) the fraction of the budget spent on the environmental services in fixed supply is smaller than for produced goods in general, then the relative "price" or value of the environmental services in fixed supply will increase over time relative to the price of the produced goods at those levels of use short of the point at which congestion externalities occur.<sup>15</sup> Additionally, changes in consumer preferences clearly can affect both the quantity and the price shift parameters.

Suppose, now, that the demand for wilderness recreation in the Hells Canyon is expanding at some rate in the quantity dimension, due perhaps to changes in population and preferences, and at some other rate in the price dimension, due perhaps to changes in income and technology. Then total benefits will be increasing at a rate equal to the sum of these rates, assuming a linear imputed demand function.<sup>16</sup>

<sup>15</sup> These conditions can be derived from the Hicks-Allen two-good general equilibrium model. Details can be furnished by the authors on request.

<sup>16</sup> Let  $P_t$  = the vertical intercept at time  $t$

$Q_t$  = the horizontal intercept at time  $t$

$r_p$  = rate of growth in willingness to pay (vertical shift)

$\gamma$  = rate of growth in quantity (horizontal shift)

$B_t$  = benefits at time  $t$

Then

<sup>14</sup> For an illustration of the rapid growth in wilderness recreation, see the figures for National Forest wilderness, wild and primitive area recreation, reported by Irving Hoch.

It is easily seen that as benefits from preservation increase relative to benefits from development, the optimal short-run level of development  $D^*(t)$  decreases.<sup>17</sup>

We can now show, in the analytical framework of the preceding section, the effect of this trend on optimal policy. If  $D^*(t)$  is monotone decreasing, then there is in effect an infinite blocked interval. Development is either frozen at the initial level  $D(0)$ , or jumps, at  $t=0$ , to some higher level  $\bar{D}$ , and is then frozen. If there is some initial investment, obviously  $q(0)=0$ . Also,  $\lim_{t \rightarrow \infty} q(t)=0$ , because

$$(21) \quad -e^{-\rho t} \leq q(t) = \sigma p(t) - e^{-\rho t} \leq 0,$$

$$(a) \quad \begin{aligned} B_t &= (1/2)P_t Q_t \\ &= 1/2 (P_0 e^{r_y t}) (Q_0 e^{\gamma t}) \\ &= 1/2 P_0 Q_0 e^{(r_y + \gamma)t} \end{aligned}$$

The increment over an infinitesimal period is

$$(b) \quad \frac{dB_t}{dt} = 1/2 P_0 Q_0 e^{(r_y + \gamma)t} (r_y + \gamma),$$

and the percent rate of increase is

$$(c) \quad \frac{\frac{dB_t}{dt}}{B_t} = \frac{1/2 P_0 Q_0 e^{(r_y + \gamma)t} (r_y + \gamma)}{1/2 P_0 Q_0 e^{(r_y + \gamma)t}} = r_y + \gamma$$

<sup>17</sup> Ignoring investment, total benefits at any time  $t$  from an area of size  $L$ , where  $L=P+D$ , and benefits  $B^P$  from the preserved area  $P$  are increasing relative to benefits  $B^D$  from developed area  $D$  at a rate of  $\alpha'$ , are

$$(a) \quad \begin{aligned} B &= B^P(P, t)e^{\alpha' t} + B^D(D, t) \\ &= B^P(L - D, t)e^{\alpha' t} + B^D(D, t) \end{aligned}$$

Optimal  $D$ ,  $D^*$ , is found by differentiating with respect to  $D$  and setting equal to zero.

$$(b) \quad \frac{\partial B}{\partial D} = -B_P^P e^{\alpha' t} + B_D^D = 0$$

or

$$B_D^D = B_P^P e^{\alpha' t}$$

As  $t$  increases,  $e^{\alpha' t}$  increases, so that  $B_P^P$  (the marginal benefits of preservation) must be decreasing, implying that  $P^*$  is increasing—and  $D^*$  decreasing.

and

$$(22) \quad \lim_{t \rightarrow \infty} e^{-\rho t} = 0$$

On the blocked interval  $(0, \infty)$  then,  $D(t) = \bar{D}$ , with

$$(23) \quad D(0) \leq \bar{D},$$

$$(24) \quad \int_0^\infty e^{-\rho t} r(\bar{D}, t) dt \geq 0$$

(but the strict inequality cannot hold in both) and

$$(25) \quad \int_t^\infty e^{-\rho t} r(\bar{D}, t) dt > 0 \quad 0 < t < \infty$$

For the projected development in the Hells Canyon, the interpretation of the analytical results is that it should be undertaken immediately, if at all. In symbols, if

$$(26) \quad \begin{aligned} \int_0^\infty [B^P(L - \bar{D}, t) + B^D(\bar{D}, t) \\ - I(t)] e^{-\rho t} dt > \int_0^\infty [B^P(L - D(0), t) \\ - I(t)] e^{-\rho t} dt, \end{aligned}$$

where  $\bar{D} > D(0)$ , then some initial development, to a level of  $\bar{D}$ , will be optimal. If the inequality is reversed, then no further development beyond  $D(0)$  should be undertaken. In the next section, a partial and approximate evaluation of these present value integrals is attempted, with  $\bar{D}$  corresponding to the most profitable level of development, the High Mountain Sheep project.

Before proceeding with the evaluation, a few qualifying remarks about the analytical results may be made. First, although a particular program, in this case nondevelopment, may be indicated given current anticipations, it can be revised (in the direction of further development) at any time following the emergence of new and unanticipated relationships in the econ-

omy, as for example, a reversal of the historic decline in energy costs. Or, though a particular level of development, corresponding say to High Mountain Sheep in the Hells Canyon, may be optimal for the purposes of power generation, a more intensive level may be indicated by the inclusion of another purpose, for example, flat water recreation. In fact, this is not now true for development in the Hells Canyon, because the separable costs of high density recreation facilities would exceed their benefits.<sup>18</sup>

Second, the somewhat abstract nature of the development measure  $D$  might be noted.  $D$  can increase, for example, by developing additional sites along the river, the construction of facilities to accommodate larger numbers of flat water recreation seekers, the penetration by roads of virgin sections, etc.

Third, to what extent, if any, has the case for preservation been overstated by the absolute restriction on reversibility, and can the restriction be relaxed? Our view, as stated earlier, is that the irreversibility of development is fundamental to the problem. This does not, however, mean that it must be absolute. Two kinds of reversal are possible, or at least conceivable. One is the restoration of an area by a program of direct investment. This would seem to have little relevance, however, for the sorts of phenomena with which we are mainly concerned: an extinct species or ecological community that cannot be resurrected, a flooded canyon that cannot be replicated, an old-growth redwood forest that cannot be restored, etc.<sup>19</sup>

The other kind of reversal is the natural reversion to the wild, which, though also

seemingly of little relevance to our main concerns, is easily fit into the analytical framework. Suppose some (constant, though this is not necessary) nonzero rate of reversion,  $\delta$ . Then  $D'(t) = D(t) e^{-\delta t}$ , where  $D'(t)$  is development subject to reversion. It is not clear how much additional flexibility this gives to investment policy. Even in situations in which  $\delta$  is significantly different from zero, it may be much smaller than the desired rate of decrease as determined by changing technology and demand and unconstrained by nature.

### III

In this section we present estimates of the costs and benefits associated with the alternatives for Hells Canyon. There are various services which the canyon can provide if preserved in its natural state. The value of some have become measurable through recent advances in economic analysis, for example, outdoor recreation, while the value of others are still intractable to economic measurement, for example, preservation of rare scientific research materials. Since we cannot measure the benefits in toto, we ask, rather, what would the present value of preserving the area need to be to equal or exceed the present value of the developmental alternative. Owing to the inverse relationship between  $\pi$  and  $\alpha_t$  (see below) the initial year's preservation benefit may need to be only very modest in comparison with the development benefit. This is illustrated in simplified, discrete form in equation (27) below.

$$(27) \quad b_p^m = \sum_{t=1}^T \frac{b_0/(1+\pi)^t}{(1+i)^t} \div \sum_{t=1}^{T'} \frac{\$1(1+\alpha_t)^t}{(1+i)^t}$$

where:

$b_p^m$  = the minimum initial year's benefit

<sup>18</sup> See testimony of Krutilla, *FPC hearings*, R-5840 and R-6494-6499.

<sup>19</sup> This is not to deny its relevance in some contexts, as shown for example in the clean-up and revegetation of certain former coal mining areas.

required to make the present value of benefits from preserving the area equal to the present value of the development benefits,

$b_0$  = the initial year's development benefit,

$\pi$  = the simplified representation of technological change for the development alternative and is defined in the Appendix

$T$  = the relevant terminal year for the development alternative,

$T'$  = the relevant terminal year for the preservation alternative,

$i$  = the discount rate,

$\alpha_t$  = the percent rate of growth in annual benefits as described in footnote 16.

This is the required initial year's benefit from preservation which makes the two alternatives equivalent and relation (26) an equality.

The terminal years for each choice,  $T$  and  $T'$ , are determined by the years in which the discounted annual benefit falls to zero. They need not and probably would not be the same. Any change in the relative annual values of the incompatible alternatives would result in different relevant time horizons.<sup>20</sup> For convenience in computation, we select  $T$  and  $T'$  as the years in which the increment to the present value of net benefits of each alternative falls to \$0.01 per \$1.00 of initial year's benefits.

Now  $\pi$  in the numerator of equation (27) is derived from our technical change model (see the Appendix). The value of  $\pi$  depends on a) investment per unit of thermal capacity, b) cost per kilowatt hour of thermal energy, and c) the rate of advance in technical efficiency. We have relied on construction cost data provided by Fed-

eral Power Commission (*FPC*) staff witnesses;<sup>21</sup> taken energy costs to be increasing from 0.98 mills per kilowatt hour in the early stage to 1.28 mills per kilowatt hour in the later period of analysis owing to projected increases in cost of processing nuclear materials;<sup>22</sup> and selected rates of technological progress of 3 to 5 percent, believed to bracket the relevant range.<sup>23</sup> Using such data in our technological change model we find that gross hydroelectric benefits will be overstated between 5 and 11 percent when technological change is not introduced into the analysis.<sup>24</sup> While the difference in gross benefits may not be very large, if the two alternatives are close cost competitors, such small differences can make a large difference in net benefits. In short, using a medium value for all of the parameters tested results in a reduction in the net present value by approximately a half in the Hells Canyon hydroelectric evaluation.<sup>25</sup>

<sup>21</sup> See testimony of *FPC* witness Joseph J. A. Jessell, *FPC* hearings, and Exhibit No. R-54-B.

<sup>22</sup> See testimony of *FPC* witness I. Paul Chavez, "In the Matter of . . .," and Exhibit No. R-107-B.

<sup>23</sup> The rate of technological change was computed from data presented in the biennial reports of *Electrical World* over a period representing a consistent method of reporting, 1950-68. It must be acknowledged that the model used for computational purposes is applicable to the period of the past, dominated by use of fossil fuels and not specifically relevant to the yet unspecified changes in technology of the future, doubtless to be tied closely to nuclear reactors. The argument, however, is that while the relevant models would differ, the effects of technological change on costs of generation will be of the same or greater order of magnitude and should not be ignored. (See testimony of Krutilla, *FPC* hearings, R-5838.) Although, as noted earlier, at least some of the reduction in costs may be balanced by a rise in environmental pollution from the more efficient fossil fuel plants, estimated costs of dealing with the thermal pollution from a nuclear plant are included in our calculations (though not the possible but unknown costs of radioactive waste disposal).

<sup>24</sup> See Table 1, Exhibit R-670, *FPC* hearings, for the complete range of values resulting from the computational model given in the Appendix.

<sup>25</sup> See testimony of John V. Krutilla, *FPC* hearings, R-5842-3 and Exhibit Nos. R-669 and R-671.

<sup>20</sup> Since control theory has not previously been applied in public sector benefit-cost studies, the time horizons have been selected arbitrarily.

In our discussion of benefits from preservation in the last section, especially in footnote 16, we took  $\alpha_i$  to be constant. This is plausible, however, only so long as the capacity of the area for recreation activity is not reached. If demand for the wilderness recreation services of the area is growing, congestion externalities eventually will arise. That is, a point will be reached beyond which use of the area by one more individual per unit time will result in a diminution of the utility obtained by others using the area. For purposes of this analysis, this point is taken as the "carrying capacity" of the area. If the benefits of additional use exceeded the congestion costs, total benefits could be increased by relaxing this constraint.<sup>26</sup> But we seek here to define a quantity of constant quality services, the value of which will be a lower bound for preservation benefits. Counting from the base year, let  $k$  be the year in which use of the area reaches capacity,  $m$  the year in which  $\gamma$  falls to the rate of growth of population, and  $d$  the rate of decay of  $\gamma$ .

Beyond some point, then, annual benefits do not grow at a uniform rate over time but depend upon the values taken by  $\gamma$ ,  $r_v$ ,  $k$ ,  $d$ , and  $m$ . The particular values taken, i.e.,  $\gamma$  of 10 percent and  $k$  of 20 years, with alternative assumptions for purposes of sensitivity analyses, were chosen for reasons given elsewhere.<sup>27</sup> A discount rate of 9 percent with alternatives of 8 and 10 percent was the result of independent study.<sup>28</sup> The selection of the value of  $m$  for 50 years, with alternative assumptions of 40 and 60, was governed by both the rate of growth of general demand for wilderness or primitive area recreation, and the estimated "saturation

level" for such recreational participation for the population as a whole. Finally, the range of values for  $r_v$  was taken from what we know about the conventional income elasticity of demand<sup>29</sup> as related to the special case of a unique resource in fixed supply and growth in per capita income over the past two or three decades.<sup>30</sup>

To contrast the results of our analysis with traditional benefit-cost analysis, consider the computed initial year's preservation benefit (Table 1) corresponding to  $i$  of 9 percent,  $r$  of 0.04,  $\gamma$  of 10 percent and  $k$  of 20 years,  $m$  of 50 years and  $r_v$  of 0.05; namely, \$80,122. This sum compares with the sum of \$2.9 million, which represents the "levelized" annual benefit from the hydro-electric development, when neither adjustments for technological progress have been made in hydroelectric power value computations, nor any site value (i.e., present value of opportunity benefits foreclosed by altering the present use of Hells Canyon) is imputed to costs. Typically then, the question would be raised whether or not the preservation value is equal to or greater than the \$2.9 million annual benefits from development.

Let us now consider the readily quantifiable opportunity benefits which would be foreclosed by development of the canyon. These are based on studies conducted by the Oregon and Idaho fish and game commissions in cooperation with the U.S. Forest Service and monitored by an observer representing the applicants for the *FPC* license. Presented in summarized form they appear in Table 2.<sup>31</sup>

While systematic demand studies of the several different recreational activities were not conducted in connection with the imputed values, given what is known about prices paid for fishing and hunting rights where such rights are vested in

<sup>26</sup> For a detailed discussion of this and other considerations in determining the capacity of a natural area for recreation activity, see Fisher and Krutilla.

<sup>27</sup> See testimony of John V. Krutilla, *FPC* hearings, R-5859-73.

<sup>28</sup> See Eckstein and Arnold Harberger, and also James Seagraves.

<sup>29</sup> See Cicchetti, Joseph Seneca and Paul Davidson.

<sup>30</sup> See footnote 15.

<sup>31</sup> See testimony of John V. Krutilla, *FPC* hearings, R-5877-, Table 3 R-5878-9, R-5880-4.

TABLE 1—INITIAL YEAR'S PRESERVATION BENEFITS,  $b_p^m$ , (GROWING AT THE RATE  $\alpha_t$ ) REQUIRED IN ORDER TO HAVE PRESENT VALUE EQUAL TO DEVELOPMENT

$r_v$	$\gamma=7.5$ Percent k=25 years	$\gamma=10$ Percent k=20 years	$\gamma=12.5$ Percent k=15 years
$i=8$ Percent,	m=50 years,	$r=0.04$ ,	$PVC_{1...T} = \$18,540,000$
0.04	\$138,276	\$109,149	\$106,613
0.05	87,568	70,363	70,731
0.06	48,143	39,674	41,292
$i=9$ Percent,	m=50 years,	$r=0.04$ ,	$PVC_{1...T} = \$13,809,000$
0.04	\$147,422	\$115,008	\$109,691
0.05	101,447	80,122	78,336
0.06	64,300	51,700	52,210
$i=10$ Percent,	m=50 years,	$r=0.04$ ,	$PVC_{1...T} = \$9,861,000$
0.04	\$142,335	\$110,240	\$103,030
0.05	103,626	80,888	77,232
0.06	71,369	56,397	55,194

Sources: Exhibit No. R-671, R-672, FPC hearings, and Transcript R-5869-5873.

Where:

$i$  = discount rate

$r_v$  = annual rate of growth in price for a given quantity

$\gamma$  = annual rate of growth of quantity demanded at given price

k = number of years following initial year upon which carrying capacity constraint becomes effective

m = number of years after initial year upon which  $\gamma$  falls to rate of growth of population

$PVC_{1...T}$  = present value of development (adjusted)

$r$  = annual rate of technological progress in the development case

private parties, we feel our estimates are rather conservative.<sup>32</sup>

<sup>32</sup> See also William Brown, Ajmer Singh and Emery Castle; Stephen Mathews and Gardiner Brown; and Peter Pearce for more systematic evaluation of the Oregon and Washington Steelhead-Salmon Fisheries and other big game resource values, and the estimated willingness to pay. On the basis of all the evidence available to us the imputation of values in the Hells Canyon case appear to be most conservative. It should be noted, however, that two assumptions are made in order that the values appearing in Table 2 represent net benefits, consistent with the benefits estimated for the hydro development. One assumption is that there are no adequate substitutes of like quality, i.e., other primitive scenic areas are either congested or being rationed, conditions which are widely encountered in national parks and over much of the wilderness system. Secondly, it is assumed that the demand unsatisfied by virtue of the transformation of the Hells Canyon would impinge on the margin in other sectors of the economy characterized by free entry and feasibility of augmenting supplies, i.e., incremental costs will equal incremental benefits.

Considering the estimates one might argue, for example, that the preservation benefits shown are roughly only a third (\$.9 to \$2.9 million) as large as would be

TABLE 2—ESTIMATED INITIAL YEAR'S QUANTIFIABLE PRESERVATION BENEFITS

Recreation Activity	Visitor Days		Imputed Benefits
	Initial Year	Unit Value	
Streamside Use			
Angling	84,000	\$ 5.00	\$420,000
Canyon Area Hunting			
Big game	7,000	25.00	175,000
Upland bird	1,000	10.00	10,000
Increased value of remaining hunting experience	29,000	10.00	290,000
Total Quantifiable Opportunity Benefits			\$895,000

required in comparisons based on traditional analysis of similar cases. By introducing the differential incidence of technological progress on, and growth in demand for, the mutually exclusive alternative uses of the Hells Canyon, we reach quite a different conclusion. The initial year's preservation benefit, subject to reevaluation on the basis of sensitivity tests, appears to be an order of magnitude (\$900,000 to \$80,000) larger than it needs to be to have a present value equaling or exceeding the present value of the development alternative. Thus we get results significantly different from traditional analysis.

What about the sensitivity of these conclusions to the particular values the variables used in our two simulation models are given? Sensitivity tests can be performed with the data contained in Table 1, along with additional information available from computer runs performed. Some of these checks are displayed in Table 3.

TABLE 3—SENSITIVITY OF ESTIMATED INITIAL YEAR'S REQUIRED PRESERVATION BENEFITS TO CHANGES IN VALUE OF VARIABLES AND PARAMETERS (at  $i=9$  percent)

Variable	Variation in Variable		Percent Change in Preservation Benefit	
	From	To	Change	
$r_d$	0.04	0.05	25	39-49
$r$	0.04	0.05	25	25
$k^a$ (years)	20	25	25	30-40
$\gamma$ (percent)	10	12.5	25	-4 to +7
$m$ (years)	40	50	25	3

<sup>a</sup> The 25 percent change in years before carrying capacity is reached translates into a 40 percent change in carrying capacity at the growth rate of 10 percent used here.

Given the estimated user days and imputed value per user day, the conclusion regarding the relative economic merit of the two alternatives is not sensitive within a reasonable range, to the particular values chosen for the variables and pa-

rameters used in the computation models.

There is need, however, for another set of tests when geometric growth rates are being used. We might regard these as "plausibility analyses." For example, the ratio of the implicit price to the projected per capita income in the terminal year was examined and found to equal  $2.5 \times 10^{-3}$ . At today's per capita income level this is comparable to a user fee of approximately \$10.00. Similarly, the ratio of the terminal year's preservation benefit to the GNP in the terminal year is found to be  $4.0 \times 10^{-7}$ . This value compares with a ratio of the total revenue of the applicants in 1968 to GNP of  $5.0 \times 10^{-4}$ . The year at which the growth rate in quantity of wilderness-type outdoor recreation services demanded falls to the rate of growth of the population must also be checked to ensure that the implicit population participation rate is something one would regard as plausible. Such tests were performed in order to avoid problems which otherwise would stem from use of unbounded estimates. We found our assumed initial rate of 10 percent, appropriately damped over time, was a realistic value.

Finally, since the readily observed initial year's benefits appear to be in excess of the minimum which would be required to have their present value exceed the present value of development, the computation is concluded at this point. Note, however, that since the analysis relies implicitly on the price compensating measure of consumer surplus, the resulting estimate of preservation value would be for this reason, as well as the restricted carrying capacity, a lower bound. Moreover, in seeking maximum expected benefits, we have implicitly assumed a neutral attitude toward risk. In fact, some preliminary findings as to the effect of uncertainty on optimal environmental policy suggest that there may be a kind of risk premium, or other adjustment, in the direction of re-

ducing benefits from development relative to preservation.<sup>33</sup>

#### IV

In Section I of this study we propose a model for the allocation of natural environments between preservation and development, and show that it will in general be optimal to refrain from some development indicated by current benefits and costs if, in the relatively near future, "undevelopment," which is impossible, would be indicated. In Section II we show that if, as in the case of the proposed development in the Hells Canyon, benefits from development are decreasing over time relative to benefits from preservation, it will be optimal to proceed with the development immediately, if at all.<sup>34</sup> In Section III we consider this question in detail for the case of the Hells Canyon, and show that it will not, in fact, be optimal to undertake even the most profitable development project there. Rather the area is likely to yield greater benefits if left in its natural state.

#### APPENDIX

Over the first 30-year period, taken as the useful life of a thermal facility, let  $PVC_t$  represent the present value of annual costs per kilowatt of the thermal alternative in year  $t$ :

$$PVC_1 = C_1 + E(8760F)$$

$$PVC_2 = \left\{ C_1 + [E8760(F - k)] + \frac{E}{(1+r)} (8760k) \right\} \left( \frac{1}{(1+i)} \right)$$

⋮

<sup>33</sup> See Cicchetti and A. Myrick Freeman, and Fisher. This is an important question and one which bears on the design of optimal policies for pollution control as well, but further consideration is beyond the scope of this paper.

<sup>34</sup> This is consistent with the obvious differences in views held by members of affluent societies and less

$$PVC_n = \left\{ C_1 + E[8760(F - (n-1)k)] + \frac{E}{(1+r)^{n-1}} [8760(n-1)k] \right\} \cdot \left( \frac{1}{1+i} \right)^{n-1} \quad \text{for } 1 < n < 30$$

where

$C_1$  = Capacity Cost/KW/yr during first 30-year period

$E$  = Energy Cost/KWh

$F$  = The plant factor; (.90)

$k$  = a constant representing the time decay of the plant factor (.03)

$i$  = the discount rate

$r$  = the annual rate of technological progress

Writing out the  $n$ th term yields:

$$PVC_n = \frac{C_1}{(1+i)^{n-1}} + \frac{8760EF}{(1+i)^{n-1}} - \frac{8760Ek(n-1)}{(1+i)^{n-1}} + \frac{8760Ek(n-1)}{[(1+r)(1+i)]^{n-1}}$$

These terms can be summed individually using standard formulas for geometric progressions and then factored to form:

$$PVC_1, \dots, 30 = \sum_{n=1}^{30} PVC_n = (C_1 + 8760EF) \cdot \left[ \frac{1 - a^{30}}{1 - a} \right] - \frac{8760Ek}{i} \cdot \left\{ \frac{1 - a^{29}}{1 - a} - 29a^{29} \right\} + \frac{8760Ek}{(1+r)(1+i) - 1} \left\{ \frac{1 - b^{29}}{1 - b} - 29b^{29} \right\}$$

where

$$a = \left( \frac{1}{1+i} \right)$$

$$b = \frac{1}{(1+r)(1+i)}$$

developed countries on these and related environmental issues.

Over years 31, . . . , T the cost expressions are similar except that we are dealing with only T-30 additional years and all terms thus get discounted by a factor of  $(1/1+i)^{30}$ . Hence, using similar formulas for the sum of geometric series the present value of annual costs per kilowatt from this later period is determined to be:

$$PVC_{31, \dots, T} = \sum_{n=31}^T PVC_n = \left( \frac{1}{1+i} \right)^{30} \cdot \left\{ C_{II} + 8760E'F \left[ \frac{1 - a^{(T-30)}}{1 - a} \right] - \frac{8760E'k}{i} \left[ \frac{1 - a^{(T-31)}}{1 - a} - 19a^{(T-31)} \right] + \frac{8760E'k}{(1+r)(1+i) - 1} \left[ \frac{1 - b^{(T-31)}}{1 - b} - 19b^{(T-31)} \right] \right\}$$

where

$$C_{II} = \frac{C_I}{(1+r)^{30}}$$

$$E' = \frac{E}{(1+r)^{30}}$$

The overall present value is:

$$PVC_{1, \dots, T} = PVC_1 + \dots + PVC_{30} + PVC_{31} + \dots + PVC_T$$

Traditional analyses are based essentially on the model given below.

$$K = \sum_{n=1}^T \frac{[C_I + E(8760F)]}{(1+i)^{n-1}}$$

or, which is equivalent,

$$= [C_I + E(8760F)] \left[ \frac{1 - a^T}{1 - a} \right]$$

We can therefore determine a relationship between the traditional measure of development benefit (K) and the measure outlined in this appendix ( $PVC_{1, \dots, T}$ ) in order to

define the simplified measure of technological change ( $\pi$ ) utilized in the text above.

$$\frac{K}{PVC_{1, \dots, T}} = \frac{\sum_{t=1}^T \frac{b_0}{(1+i)^t}}{\sum_{t=1}^T \frac{b_0/(1+\pi)^t}{(1+i)^t}} = T / \sum_{t=1}^T \frac{1}{(1+\pi)^t}$$

## REFERENCES

- K. J. Arrow, "Optimal Capital Policy with Irreversible Investment," in J. N. Wolfe, ed., *Value, Capital and Growth*, Chicago 1968, pp. 1-20.
- and M. Kurz, "Optimal Growth with Irreversible Investment in a Ramsey Model," *Econometrica*, Mar. 1970, 38, 331-44.
- and A. C. Fisher, "Environmental Preservation, Uncertainty, and Irreversibility," *Quart. J. Econ.*, forthcoming.
- H. J. Barnett and C. Morse, *Scarcity and Growth*, Baltimore 1963.
- W. G. Brown, A. Singh, and E. N. Castle, "Net Economic Value of the Oregon Salmon-Steelhead Sport Fishery," *J. Wildlife Manage.*, Apr. 1965, 29, 66-79.
- O. R. Burt and R. Cummings, "Production and Investment in Natural Resource Industries," *Amer. Econ. Rev.*, Sept. 1970, 60, 576-90.
- C. J. Cicchetti, J. Seneca, and P. Davidson, *The Demand and Supply of Outdoor Recreation*, Washington 1969.
- C. J. Cicchetti and A. M. Freeman, III, "Option Demand and Consumer Surplus," *Quart. J. Econ.*, Aug. 1971, 85, 528-39.
- O. Eckstein, *Water Resource Development*, Cambridge 1958.
- , in *Economic Analysis of Public Investment Decisions: Interest Rate Policy and Discounting Analysis*, Hearings, Joint Economic Committee, 90th Congress, 2d Sess., Washington 1968.
- A. C. Fisher and J. V. Krutilla, "Determina-

- tion of Optimal Capacity for Resource-Based Recreation Facilities," *Natur. Resources J.*, forthcoming.
- A. Harberger, in *Economic Analysis of Public Investment Decisions: Interest Rate Policy and Discounting Analysis*, Hearings, Joint Economic Committee, 90th Congress, 2d Sess., Washington 1968.
- J. R. Hicks and R. G. D. Allen, "A Reconsideration of the Theory of Value," Part I, II, *Economica*, Feb., May 1934, 1, 52-76; 196-219.
- I. Hoch, "Economic Analysis of Wilderness Areas," in *Wilderness and Recreation—a Report on Resources, Values and Problems*, ORRRC Study Report No. 3, Washington 1962, pp. 203-64.
- M. M. Hufschmidt, J. V. Krutilla, and J. Margolis, *Standards and Criteria for Formulating and Evaluating Federal Water Resources Development. Report to the Bureau of the Budget*, Washington 1961.
- J. V. Krutilla, "Conservation Reconsidered," *Amer. Econ. Rev.*, Sept. 1967, 57, 777-86.
- and O. Eckstein, *Multiple Purpose River Development*, Baltimore 1958.
- S. B. Mathews and G. S. Brown, *Economic Evaluation of the 1967 Sport Salmon Fisheries of Washington*, Washington Dept. of Fisheries Technical Report 2, Olympia 1970.
- R. N. McKean, *Efficiency in Government Through Systems Analysis*, New York 1958.
- E. J. Mishan, "The Postwar Literature on Externalities: An Interpretive Essay," *J. Econ. Lit.*, Mar. 1971, 9, 1-28.
- P. H. Pearse, "A New Approach to the Evaluation of Non-Priced Recreation Resources," *Land Econ.*, Feb. 1968, 44, 87-99.
- C. G. Plourde, "A Simple Model of Replenishable Natural Resource Exploitation," *Amer. Econ. Rev.*, June 1970, 60, 518-22.
- N. Potter and F. T. Christy, Jr., *Trends in Natural Resource Commodities*, Baltimore 1962.
- J. A. Seagraves, "More on the Social Rate of Discount," *Quart. J. Econ.*, Aug. 1970, 84, 430-50.
- V. L. Smith, "Economics of Production from Natural Resources," *Amer. Econ. Rev.*, June 1968, 58, 409-31.
- P. O. Steiner, "The Role of Alternative Cost in Project Design and Selection," *Quart. J. Econ.*, Aug. 1965, 79, 417-30.
- H. Wold and L. Jureen, *Demand Analysis*, New York 1953.
- Federal Power Commission (FPC), "In the Matter of: Pacific Northwest Power Company and Washington Public Power Supply System," hearings, Washington 1970.
- Policies, Standards, and Procedures in the Formulation, Evaluation, and Review of Plans for Use and Development of Water and Related Land Resources*, prepared under the direction of the President's Water Resources Council, Senate Document no. 97, 87th Cong., 2d sess., Washington 1964.
- , Supp. no. 1, "Evaluation Standards for Primary Outdoor Recreation Benefits," Ad Hoc Resources Council, Washington 1964.
- Proposed Practices for Economic Analysis of River Basin Projects*, report to the Inter-Agency Committee on Water Resources, prepared by the Subcommittee on Evaluation Standards, Washington 1958.
- "Steam Station Cost Survey," *Electrical World*, Nov. 3, 1969.

# The Preventive Tariff and the Dual in Linear Programming

By LEONARD WAVERMAN\*

Traditional discussions of the effects of tariffs consider a sector such as manufacturing or an industry (or an entire economy) on an aggregate basis, neglecting the geographic dispersion of production and consumption which is in fact the case. The preventive tariff calculated by considering all production and consumption as taking place at one point is, therefore, some mean of the various tariffs which would be in effect at the various different locations of markets. For goods where transportation costs are low it would be impossible to maintain differentials in import duties for different regions within a country since interregional flows do not face tariffs, all imports would enter at the region with the lowest duty and then be shipped within the country to other consuming areas. Where transportation costs are a significant portion of the final delivered price, it is feasible to maintain a system of differential duties on the same product, if the difference between the tariffs at regional entry points is less than the cost of reshipping the commodity within the country.

Of course, such a system of differential tariffs would be very difficult politically to introduce, even where differential tariffs might make good economic sense. It is

shown below, that since a uniform tariff leads to various degrees of protection in domestic markets, a tariff policy which is aimed at assisting domestic producers in one particular market yields excess profits in other markets.

The primary purpose of this paper is to suggest a framework for the analysis of tariff policy in a geographic context for goods where transportation costs are a significant portion of the market price.<sup>1</sup> The analysis will be applied to the market for natural gas in North America. A simple linear programming transportation model is solved for the dual values at several Canadian markets. It is shown that the change in dual values between free trade and constrained trade solutions can be interpreted as the minimum tariff necessary to prevent entry. These preventive tariffs<sup>2</sup> vary widely among markets and depend on the location of the supplier in the free trade solution. Finally, the paper attempts to analyze whether the actual tariff itself was the major obstacle to an increased penetration of Canadian gas markets by American suppliers.

Between 1953 and January 1, 1969, a tariff of three cents per thousand cubic feet (*MCF*) was imposed by the Canadian

\* Assistant professor of economics, University of Toronto. This paper is an extension of my doctoral thesis entitled "Natural Gas Flows: Costs of a Border" submitted in August 1969 to the department of economics at the Massachusetts Institute of Technology. Morris Adelman, Paul MacAvoy, and Frank Fisher were most helpful in directing the thesis. Resources for the Future, Inc. aided research through a fellowship grant. The comments of George Borts and several anonymous referees were instrumental in improving the paper.

<sup>1</sup> The impact of transport costs on location of production and on the comparability of origin and destination taxes has been studied by Walter Isard and Merton Peck, and W. M. McNie among others. However, these studies do not consider the dispersion of consumption and production, i.e. the importance of transport costs on intracountry movements. As a result these studies do not consider the need for different tariffs in different markets within one country.

<sup>2</sup> Examples of the literature on preventive and optimal tariffs are Harry Johnson, James Melvin and Bruce Wilkinson, and A. H. H. Tan.

federal government on all imports of natural gas into Canada. The tariff was but one of many factors tending to limit the penetration of the Canadian market by American natural gas suppliers. Other Canadian government policies included the refusal of export and import licenses,<sup>3</sup> and the promotion of a transcontinental East-West pipeline completed in 1956. The American government also acted so as to limit the international trade in natural gas.<sup>4</sup>

A linear programming model (called the free trade solution) is used to estimate the hypothetical flow pattern which would occur in the absence of any trade restrictions. The solution to a second linear programming model (called the constrained model) is then estimated, where constraints are changed in the free trade model so that no foreign supplies penetrate the domestic Canadian market. As is shown in a later section the difference in the dual variables for the domestic markets (shadow delivered prices) between the free trade and the constrained solutions represents a measure of the cost disadvantages of domestic suppliers as compared to foreign suppliers in these markets. The necessary preemptive tariff is therefore equal to this comparative disadvantage of domestic suppliers in the home market.

## I. The Models

### *Free Trade Solution*

In the absence of any constraints on trans-border flows, one can estimate the hypothetical distribution of natural gas flows in North America by minimizing the costs to the final consumer subject to the constraints that demand in each market be at least met (constraint (2)),

<sup>3</sup> See Waverman (1972a), ch. 2.

<sup>4</sup> For a set of divergent views as to the reasons for this policy see Hugh Aitken and the Royal Commission on Energy *First Report*.

and that no field ship more than its capacity (constraint (3)).

Minimize

$$(1) \quad Z = \sum_{i=1}^m \sum_{j=1}^n [(ad_{ij}g_{ij} + p_i)X_{ij}]$$

subject to

$$(2) \quad \sum_{i=1}^m X_{ij} \geq D_j \quad (j = 1, \dots, n)$$

$$(3) \quad \sum_{j=1}^n X_{ij} \leq S_i \quad (i = 1, \dots, m)^5$$

where

$X_{ij}$  = the flow of gas from field  $i$  to market  $j$  in thousand cubic feet per day (MCFD);

$a$  = the cost of shipping one unit of gas MCFD one mile;

$d_{ij}$  = the distance in miles between field  $i$  and market  $j$  (around obstacles which a pipeline could not cross);

$g_{ij}$  = the terrain factor for route  $ij$  which adjusts the unit costs  $a$  for the particular terrain through which the pipeline passes;

$p_i$  = the cost of producing one unit of gas MCFD at the field.

By inserting appropriate estimates of  $a$ ,  $d$ ,  $g$ ,  $p$ ,  $D$ ,  $S$ , the flows  $X_{ij}$  are determined which minimize the objective function (1)

<sup>5</sup> It is implicit that  $\sum^n D_j \leq \sum^m S_i$ . It is obvious that if  $\sum^n D_j > \sum^m S_i$ , constraint (2), that demand be met in each market, could not be fulfilled. If  $\sum^n D_j = \sum^m S_i$ , then only  $n+m-1$  equations are independent. Suppose all  $n$  demand constraints are met, as are  $m-1$  supply constraints, then the  $m$ th supply constraint must also be met. Since the number of dual variables equals the number of constraints in the primal,  $m+n$  dual values would be determined by  $m+n-1$  independent equations. No unique set of dual prices could be determined. As a result, some excess capacity at a field is required to find a unique solution for both the primal and the dual. This excess capacity is estimated by the amounts of gas flared or wasted at the supply point, gas not actually shipped to the market. In total, excess capacity is small, some 1.5 percent of total demand. It is also implicit that  $X_{ij} \geq 0$ .

subject to the demand and supply constraints.<sup>6</sup>

Estimates of transportation costs were taken from a paper by C. L. Dunn and are close to estimates given by Adelman. The set of terrain factors was obtained from the National Energy Board, Ottawa. Production costs for the United States were taken to be the average of the two Federal Power Commission (FPC) area wide ceiling rates in effect for each area<sup>7</sup>: prices for "old" contracts (those signed prior to 1961) are averaged with prices for "flowing" gas (contracts signed after 1961), since these area rate prices become the unit costs of gas to interstate pipeline companies.<sup>8</sup> For Canada, the average contract price for gas shipped in 1966 was taken to be a good estimate of the costs of production. Market demand was taken to be the actual consumption of gas in North America in 1966, distributed among nineteen demand points, five of which were in Canada. Capacity of a field was assumed to be the actual 1966 production for an area. Four fields were in Canada, each of them representing the aggregate of a province's production. Of the fourteen U.S. production points used, seven represent specific fields and seven the total of a state's production. One supply point represented imports from Mexico.

The primal free trade model (equations (1)–(3)) was solved using the data de-

scribed above. Since no constraints (tariffs, export prohibitions, etc.) are placed on trans-border shipments, the solution to this model represents the flow pattern which would exist under free trade. This free trade solution is distinctly different from the actual shipments of 1966. Canada in this free trade solution exports 68 percent of its production as compared to 46 percent actually. In 1966, imports into Canada from the United States accounted for only 7 percent of Canadian consumption as compared to 41 percent in the hypothetical unconstrained solution. In particular, the flows which in the actual world move from Alberta east to Montreal do not appear in this free trade model, for in the solution eastern Canada entirely relies on U.S. producers for its gas supplies.

#### *Model-Constrained Solution*

To simulate more closely the actual movement of natural gas within North America, the free trade model is altered by substituting a set of constraints which incorporate the political realities of trade. Under present Canadian policy, western Canada must ship through to "meet demand" at eastern Canadian markets. The surplus of capacity in the Canadian west above eastern Canadian demand is then exportable.

The constrained model can be represented by:

Minimize

$$(4) \quad Z' = \sum_{i=1}^m \sum_{j=1}^n [(ad_{ij}g_{ij} + p_i)X_{ij}]$$

subject to

$$(5) \quad X_{kj} \geq D_j \quad (j = 1, \dots, h)$$

$$(6) \quad \sum_{i=1}^n X_{ij} \geq D_j \quad (j = h+1, \dots, n)$$

$$(7) \quad \sum_{j=1}^n X_{ij} \leq S_i \quad (i = 1, \dots, n)$$

<sup>6</sup> The model is given in greater detail in Waverman (1972b). For exposition purposes, this present paper also describes the models. There is some overlap between the two papers. However, the major points they make are quite separate.

<sup>7</sup> Where the courts had not settled the area rate ceiling, "in-line" rates established as an interim guide by the FPC were used.

<sup>8</sup> The averaging process consisted of determining for each area the proportion of sales in 1966 which were from reserves established prior to 1961 by assuming that the 1960 production to reserve ratio was maintained for these reserves established prior to 1961 for each year up to and including 1966. The two prices were then weighted by the proportion of 1966 sales from reserves established before and after 1960.

The objective function is identical in both these models, as is the last constraint (7), that no field ship more than its capacity. Constraint set (5) states that the shipments from the  $k$ th field (Alberta) to the markets 1 to  $h$  (eastern Canada) must be at least as great as the demand in these markets.<sup>9</sup> This constraint set is analogous to the form of regulation, Alberta is constrained to meet demand in eastern Canada.<sup>10</sup> Constraint set (6) states that the sum of flows into any one of the remaining markets (outside eastern Canada) from all the sources to that market must at least meet market demand.

Using the same data described above, the model given by equations (4) to (7) was reestimated to simulate the actual flow pattern of 1966. Tables 1A, 1B, and 1C give the transborder flows of the free trade solution, the flows across the border in the constrained solution and the differences

<sup>9</sup> The constraint (5) could read that Canadian supply met Canadian demand. However, little British Columbia gas enters the eastern Canadian market, and it was thought that the loss in precision would be offset by the gain in computation and simplification of analysis and description.

<sup>10</sup> "After the provincial regulatory boards have approved removal of gas from the respective province, the National Energy Board considers the advisability of adequate protection of requirements for the total national use, prior to allowing the natural gas to flow south to the United States." From a speech by J. W. Kerr, President, Trans Canada Pipelines Ltd. to the Institution of Gas Engineers, May 1965.

TABLE 1B—CONSTRAINED MODEL:  
TRANS-BORDER FLOWS  
(MMCFD)

	To:	U.S. #8	U.S. #10
From:			
British Columbia			222
Alberta		557	382

Notes: See Table 1C

in flows between the free trade and constrained solutions. Table 2 shows the differences between the flows of the two models and the actual shipments of 1966.

As can be observed from these tables, constraining Alberta to meet eastern Canadian demand changes seventeen flows in the network, affecting fields in Louisiana and markets as far away as California and Chicago. Forcing Alberta to ship to eastern Canadian markets must raise the total costs of meeting system demands since the flows from Alberta to eastern Canadian markets were available in the free trade solution, but were not chosen. The effect of constraining certain flows which would not appear under free trade is to introduce inefficient transportation shipments and thus penalize consumers.<sup>11</sup>

This procedure of estimating a hypo-

<sup>11</sup> Real costs increase by some 20-30 percent in eastern Canadian markets.

TABLE 1A—FREE TRADE MODEL: TRANS-BORDER FLOWS  
(MMCFD)

Ontario							
	To: N.W.	S.W.	Toronto	S.E.	Quebec	U.S. #8	U.S. #9 U.S. #10
From:							
British Columbia							222
Alberta						979	332 382
Louisiana North		187		89			
Texas Permian			243		89		
Kansas	146						

Notes: See Table 1C

TABLE 1C—DIFFERENCE IN FLOWS, CONSTRAINED MODEL AND FREE TRADE MODEL<sup>a</sup>  
(MMCFD)

To:	Ontario				Quebec	U.S. #4	U.S. #7	U.S. #8	U.S. #9
	N.W.	S.W.	Toronto	S.E.					
From:									
Alberta	146	187	243	89	89			(422)	(332)
Louisiana North		(187)		(89)		276			
Louisiana South						(422)	422		
Texas Permian			(243)		(89)		(422)		754
Wyoming								422	(422)
Kansas	(146)					146			

Notes: U.S. #4: Wisconsin, Michigan, Illinois, Indiana  
 U.S. #7: Texas, Arkansas, Mississippi, Louisiana  
 U.S. #8: Montana, Wyoming, Utah, Colorado  
 U.S. #9: New Mexico, Arizona, Nevada, California  
 U.S. #10: Washington, Oregon, Idaho

<sup>a</sup> Numbers in parentheses are negative.

thetical flow pattern in the absence of trade restrictions and comparing these flows to a simulated model of the actual flows under present government policy

rests on a number of assumptions and restrictive simplifications. First, neither the free trade nor the constrained solution assumes an existing network of pipelines.

TABLE 2—COMPARISON OF ACTUAL 1966 AND ESTIMATED FLOWS:  
FREE TRADE MODEL, CONSTRAINED MODEL  
(MMCFD)

From	Shipment	To	Actual Flows (approx- imate)	Estimated Flows	
				Free Trade Model	Constrained Model
Canada		U.S. Districts #8, #9 and #10	1,060	1,915	1,170
Alberta		Eastern Canada	632	—	754
U.S.		Eastern Canada	122	754	—
New Mexico		California	2,860	2,735	2,735
U.S. West South Central (Louisiana, Oklahoma, Texas)		U.S. #1, U.S. #2	7,300	7,960	7,960
West Virginia		U.S. #2	560	524	524
Louisiana and Mississippi		U.S. #3	2,800	3,025	2,800
U.S. West South Central		U.S. #4, U.S. #5	5,020	5,630	5,480
U.S. West South Central		U.S. #6	3,470	3,846	3,846
U.S. West South Central		U.S. #7	15,300	15,320	15,320

Source: Actual Flows: U.S.: in 1966 The Bureau of Mines discontinued the classification of the imports into a region by producing area. The figures here are for 1965 *Consumption and Production*, increased by that region's consumption growth factor between 1965 and 1966; Canada: National Energy Board *Report* (1966).

Notes: U.S. #1: Maine, Vermont, New Hampshire, Massachusetts, Rhode Island, Connecticut.

U.S. #2: New York, Pennsylvania, Ohio, Maryland, Delaware, New Jersey, D.C., Virginia, West Virginia, Kentucky.

U.S. #3: Tennessee, North Carolina, South Carolina, Alabama, Georgia, Florida.

U.S. #5: North Dakota, South Dakota, Minnesota, Iowa, Nebraska.

U.S. #6: Kansas, Missouri, Oklahoma.

Instead, the linear approximation method treats gas flows as movements which can be instantaneously changed and redirected in any direction. The constrained model simulates the actual flows in North America without actually incorporating the pipelines in place. The correspondence between the results of the constrained model and actual movements is then surprisingly close (Table 2). The free trade model reallocates actual production to actual demand had producers been able to meet demand under the assumptions of free trade and of course, the linear approximations.

There are a number of reasons why a linear approximation was used and why, more importantly, it appears to give good results. First, aggregating demand and supply to nineteen points each, results in flow levels in the model which would normally be transmitted in the largest pipes. Thus, a comparison between flow levels would not exhibit any scale economies. Secondly, these linear transportation costs are adjusted for the particular terrain through which a flow passes making the operating costs representative of the differing capital costs incurred in various geographic areas.

In addition, the model merely reallocates actual production to actual demand, yet such a reallocation leads to changes in both market and field prices. Demand and supply are not however adjusted to take account of these changes in prices.

Therefore, many approximations are used—linearity, static inelastic demand and supply. However, the results do seem reasonable. Reasonableness would seem to be a test of the value of the linear and static approximations to the real dynamic world. One other excuse can be made for the assumption of linearity. The purpose of the example is to show the value of using the dual variables in a linear program-

ming model to estimate preventive tariffs. Nonlinearity would make the dual variables extremely difficult to obtain and interpret.

## II. The Dual Solution and its Interpretation

The dual to the primal problem set out in equations (1)–(3) is:

Maximize:

$$(8) \quad Z = \sum_{j=1}^n \mu_j D_j - \sum_{i=1}^m v_i S_i$$

subject to:

$$(9) \quad \begin{aligned} \mu_j - v_i &\leq C_{ij} & i = 1, \dots, m \\ & & j = 1, \dots, n \end{aligned}$$

where

$$C_{ij} = \sum_{t=1}^m \sum_{j=1}^n [(ad_{ij}g_{ij} + p_i)X_{ij}]$$

Interpreting  $\mu_j$  as the shadow value of a unit of demand in market  $j$  (the shadow delivered price in the market) and  $v_i$  as the shadow value of a unit of field capacity in field  $i$  (the royalty rent of a field), then the objective function of the dual (8) maximizes the net receipts of suppliers subject to the constraint set (9) which states that the difference between the delivered price in market  $j$  ( $\mu_j$ ) and the royalty value of field  $i$  ( $v_i$ ) must be no greater than the actual costs of delivery from  $i$  to  $j$  ( $C_{ij}$ ).<sup>12,13</sup>

<sup>12</sup> Otherwise known as the zero profit constraint.

<sup>13</sup> The Duality Theorem (see Robert Dorfman, Paul Samuelson and Robert Solow) states that for each positive flow in the primal solution, the inequality constraint of the dual is just satisfied.

$$X_{ij} > 0 \rightarrow \mu_j - v_i = C_{ij}$$

For a zero flow level in the primal solution, the corresponding constraint in the dual is not satisfied.

$$X_{ij} = 0 \rightarrow \mu_j - v_i < C_{ij}$$

If the difference between the delivered price and the field royalty is less than the costs of shipment, then losses would be incurred if the flow were nonnegative.

There also exists a dual model corresponding to the constrained model equations (4) to (7).<sup>14</sup>

Maximize:

$$(10) \quad Z' = \sum_{j=1}^n \mu'_j D_j - \sum_{i=1}^m v'_i S_i$$

subject to:

$$(11) \quad \mu'_j - v'_i \leq C_{ij}$$

where

$$C_{ij} = \sum_i \sum_j [(ad_{ij}g_{ij} + p_i)X_{ij}]$$

In the constrained model, entry into Canada of U.S. gas was prevented by an explicit requirement that eastern Canadian markets be served by western Canadian fields. As a result of this constraint, a set of dual prices  $u'_1, \dots, u'_h$  were established for these eastern Canadian markets, as opposed to a set  $u_1, \dots, u_h$  in the free trade model. The planner who had knowledge of these dual values at the market could then set price floors  $u'_1, \dots, u'_h$  in eastern Canadian markets and achieve the same results as imposing the requirements that demand in these markets be met by Canadian production. Establishing price floors for natural gas in consuming areas would be a difficult task politically. Since prices  $u_1, \dots, u_h$  would however exist under free trade, the planner need only raise the price of foreign gas in domestic markets by the difference in dual market values  $(u'_1 - u_1), \dots, (u'_h - u_h)$  to ensure that imports would not enter.

The tariff necessary to prevent entry into domestic markets is thus equal to the difference in dual shadow prices between the free trade and constrained solutions. As is shown below, this difference  $(u'_j - u_j)$

can also be interpreted as the cost disadvantage in these markets of the domestic as compared to the foreign producer.

For two fields  $i$  and  $h$  which both deliver to a market  $j$ , the dual constraints are:

$$(12) \quad \mu_j = C_{ij} + v_i$$

$$(13) \quad \mu_j = C_{hj} + v_h$$

Therefore,

$$(14) \quad C_{ij} + v_i = C_{hj} + v_h$$

or,

$$(15) \quad v_i - v_h = C_{hj} - C_{ij}$$

The difference between the royalty values of two fields supplying the same market must equal the difference in the cost of supplying that market from the two fields. The dual variables of all the fields can be developed as combinations of the differences in the costs of supplying various markets. For example, consider a market  $f$  to which fields  $i$  and  $h$  ship but field  $h$  does not:

$$(16) \quad \mu_f = C_{if} + v_i$$

$$(17) \quad \mu_f = C_{hf} + v_h$$

$$(18) \quad \mu_f < C_{hf} + v_h$$

Therefore

$$(19) \quad C_{if} + v_i = C_{hf} + v_h$$

or

$$(20) \quad v_h - v_i = C_{if} - C_{hf}$$

but

$$v_i = C_{hj} - C_{ij} + v_h$$

Therefore

$$v_h = C_{if} - C_{hf} + C_{hj} - C_{ij} + v_h$$

or

$$v_h - v_i = (C_{if} - C_{hf}) + (C_{hj} - C_{ij})$$

The difference in the royalty values of fields  $h$  and  $i$  equals the cost advantage of field  $h$  over field  $i$  in the  $f$ th market

<sup>14</sup> Both the free trade and constrained models contain the identical number of constraints— $n$  demand constraints and  $m$  supply constraints. The dual variables for the two solutions can then be compared.

plus the cost advantage of field  $i$  over field  $h$  in the  $j$ th market.

It can also be shown that the change in the royalty values of markets between solutions can be represented by relative cost differences among fields.

In the solution to the free trade model, Alberta (field 2) does not ship to eastern Canada:

$$(21) \quad X_{2j} = 0 \quad (j = 5, \dots, 9)$$

Therefore for the corresponding dual variables:

$$(22) \quad \mu_j < C_{2j} + v_2 \quad (j = 5, \dots, 9)$$

The eastern Canadian delivered price in the unconstrained solution is not sufficient to meet the costs of shipping from Alberta.

In the solution to the constrained model, Alberta is forced to meet the demands in eastern Canadian markets. Therefore

$$(23) \quad X'_{2j} > 0 \quad (j = 5, \dots, 9)$$

and:<sup>15</sup>

$$(24) \quad \mu'_j = C_{2j} + v'_2 \quad (j = 5, \dots, 9)$$

Also consider a field ( $t$ ) which does ship to eastern Canada ( $j$ ) in the free trade solution but where no supplies are forthcoming in the constrained solution.

$$(25) \quad \mu_j = C_{tj} + v_t$$

$$(26) \quad \mu'_j < C_{tj} + v'_t$$

Therefore,

$$(27) \quad \begin{aligned} \mu'_j - \mu_j &= C_{2j} + v'_2 - C_{tj} - v_t \\ &= (C_{2j} - C_{tj}) + (v'_2 - v_t) \end{aligned}$$

but from (19) above, it is known that

$$v'_2 = C_{tk} - C_{2k} + v'_t$$

so that

$$(28) \quad \begin{aligned} \mu'_j - \mu_j &= (C_{2j} - C_{tj}) + (C_{tk} - C_{2k}) \\ &\quad + v'_t - v_t \end{aligned}$$

$v'_t$  and  $v_t$  could themselves be represented by relative cost differences among fields. While this representation of  $v'_t$  and  $v_t$  might include some further royalty value  $v_n$  or  $v'_m$ , eventually the expansion of these two terms ( $v_t$  and  $v'_t$ ) would include the royalty value of fields with excess capacity ( $v_1, v'_1 = 0$ ).

The change in delivered prices at a single market can thus be considered to represent the cost advantage of some suppliers over other suppliers in that market. In particular the change in delivered price for eastern Canadian markets between the free trade and constrained solutions represents the cost advantage of the U.S. supplier in the unconstrained solution over the Canadian supplier in the constrained solution.

The calculation of the change in shadow delivered prices for each of the markets to which Alberta ships in the constrained solution but to which U.S. fields ship in the free trade solution represents Alberta's cost disadvantage in each of these markets compared to the most efficient supplier.

A government tariff policy which is designed to prohibit American entry into eastern markets must establish tariffs which are at least as high as the cost advantage of American fields in these markets. The change in the shadow delivered price in eastern Canadian markets therefore represents the minimal tariff which could be set to prevent any entry by foreign suppliers.

Table 3 gives the differences in shadow delivered prices for the eastern Canadian markets. The minimal tariff necessary to prevent foreign entry into these geographically dispersed markets ranges from 6.2

<sup>15</sup> Note that the difference in the dual models is in the shadow prices  $v_i, \mu_j$  since demand ( $D_j$ ), supply ( $S_i$ ), and costs ( $C_{ij}$ ) are identical in the two solutions.

TABLE 3—CHANGE IN DELIVERED PRICES IN EASTERN CANADIAN MARKETS BETWEEN THE FREE TRADE AND CONSTRAINED SOLUTIONS (\$/MCFD)

Market	Shadow Free Trade	Delivered Price Constrained	Price Difference
Ontario NW (5)	.310719	.372924	.062205
Ontario SW (6)	.290358	.405088	.114730
Ontario Toronto (7)	.313260	.409091	.095831
Ontario SE (8)	.334182	.419102	.084920
Quebec (9)	.341200	.429970	.088770

cents in Ontario Northwest to 11.5 cents in Ontario Southwest. A number of rather obvious points can be drawn from this table. First, an attempt to use a single domestic import tariff to restrict imports into a specific market will yield excess profits for the shippers into other markets. For example, if the Canadian government wished to prevent entry of foreign gas into the Toronto market, then a tariff of 9.6 cents would have to be placed on all imports of gas, thus preventing entry into all eastern Canadian markets except Ontario Southwest and creating excess profits in three other markets (assuming that domestic producers could price up to the world price plus tariff). In addition, if government policy was designed to increase competition in some markets but decrease imports in another market, tariff policy could establish this goal only if the cost disadvantage of domestic producers was greater in the market where competition was needed than in the other markets. Where transportation costs and the location of producer and consumer is an important determinant of the final price, then a policy of differential tariffs for various markets is necessary, if tariffs are designed to remove the costs disadvantages of domestic producers.

Since the existing Canadian tariff of 3.0 cents per MCF was less than the cost disadvantage of Canadian producers in each

of the five domestic markets, the actual tariff cannot explain the absence of U.S. gas in Canadian markets.<sup>16</sup> Other Canadian policies, such as the refusal of import licenses rather than explicit tariff policy, prevented the flow of U.S. gas into Canadian markets.<sup>17</sup> A tariff policy is more efficient than such nonmarket constraints. Import licenses, quotas, etc. limit entry without reference to the cost disadvantages of domestic producers. There is little incentive therefore for domestic producers to decrease their relative cost disadvantage in domestic markets and this cost disadvantage may rise to a level which the public would prefer not to bear. If there are noneconomic political and social benefits to having domestic producers supply the domestic market, then a rational policy would be for the government to decide what costs the country could bear for this suboptimal economic distribution and establish differential tariffs based on the comparative disadvantage of domestic producers to limit foreign penetration of domestic markets. If the cost disadvantages of domestic producers rose, then increased penetration of domestic markets would occur so that the total costs of promoting domestic production for domestic needs would remain constant.

### III. Summary

The introduction of the geographic dispersion of suppliers and markets for a good whose transportation costs are an

<sup>16</sup> The Canadian transcontinental pipeline was constructed in 1956 not 1966. Perhaps the cost disadvantages of Canadian producers were at most 3.0 cents in 1956 but increased over the decade. A study of the changes in production costs in American fields between 1956 and 1966 showed that the cost disadvantages of Canadian producers could not have been as low as 3.0 cents per MCF in 1956 (see Waverman (1972a, ch. 7)).

<sup>17</sup> For example, in 1958 the Canadian government refused an import license to Consumers Gas of Ontario and a proposed pipeline across the unnavigable Niagara chasm under the authority of the Navigable Waters Protection Act.

important ingredient of the final price complicates tariff policy. A single domestic tariff will generate either insufficient protection or excess profits in some markets, because the cost disadvantage of domestic producers varies widely among markets. While a tariff policy is more efficient than nonmarket constraints on imports, differential tariffs should be set rather than a uniform tariff. A good method of calculating the necessary tariff in any one market is to calculate the change in dual variables in the domestic markets to be protected, when in a linear model of distribution flows, foreign free trade suppliers are replaced by domestic suppliers.

In the example given—the North American natural gas industry—the actual Canadian uniform tariff was insufficient to compensate for the comparative disadvantage of domestic suppliers.<sup>18</sup> Nonmarket constraints limited U.S. penetration of Canadian markets.

#### REFERENCES

- M. A. Adelman, "The Supply and Price of Natural Gas," *J. Ind. Econ.* supp. 1962.
- H. G. J. Aitken, "The Midwestern Case: Canadian Gas and the Federal Power Commission," *Can. J. Econ.*, May 1959, 25, 129–43.
- R. Dorfman, P. A. Samuelson and R. M. Solow, *Linear Programming and Economic Analysis*, New York 1950.
- C. L. Dunn, "The Economics of Gas Transmission," paper presented at Conference of Independent Natural Gas Association of America (INGAA), Oct. 1959.
- J. M. Henderson, *The Efficiency of the Coal Industry: An Application of Linear Programming*, Cambridge 1960.
- W. M. McNie, "The Origin Principle and Transport Costs," in C. S. Shoup, ed., *Fiscal Harmonization in Common Markets*, Vol. 1, New York 1967, pp. 265–310.
- H. G. Johnson, "The Standard Theory of Tariffs," *Can. J. Econ.* Aug. 1969, 35, 333–52.
- W. Isard and M. J. Peck, "Location Theory and International and Interregional Trade Theory," *Quart. J. Econ.*, Feb. 1954, 68, 97–114.
- J. R. Melvin, and B. W. Wilkinson, "Effective Protection in the Canadian Economy," Economic Council of Canada, *Special Study No. 9*, 1968.
- A. H. H. Tan, "Differential Tariffs, Negative Value-Added and the Theory of Effective Protection," *Amer. Econ. Rev.*, Mar. 1970, 60, 107–16.
- L. Waverman, (1972a) *Natural Gas and National Policy: A Linear Programming Model of Natural Gas Flows in North America*, Toronto 1972.
- , (1972b) "National Policy and Natural Gas: The Cost of a Border," *Can. J. Econ.*, Aug. 1972, forthcoming.
- National Energy Board, *Report*, Ottawa 1965, 1966.
- Royal Commission on Energy, *First Report*, Ottawa 1957.
- U.S. Bureau of Mines, *Consumption and Production*, Washington 1965, 1966.

<sup>18</sup> It is therefore not surprising that the tariff was eliminated in the Kennedy round of tariff concessions.

# A Choice-Theoretic Model of an Income-Investment Accelerator

By HERSCHEL I. GROSSMAN\*

The conventional income-accelerator theory of investment demand asserts that investment demand is an increasing function of changes in the level of output. This theory is frequently subjected to the warranted criticism that it is a mechanistic rather than a choice-theoretic approach to economic behavior. In contrast to the conventional accelerator theory, so-called neoclassical investment-demand theory is founded on choice-theoretic analysis, and, consequently, is presented as a preferable alternative to the accelerator theory. Moreover, since the conventional neoclassical model generates no direct dependence of investment demand upon changes in output, it is alleged that such a structural relationship is inconsistent with a choice-theoretic framework.

The essential objection to the accelerator theory involves its implications that, from the standpoint of the firm, output is an exogenous variable and that the firm simply employs sufficient inputs to produce this given level of output. The treatment of output as an exogenous datum to which the firm passively adjusts is apparently inconsistent with profit-maximizing behavior, as it is developed in the standard neoclassical model.

The purpose of the present paper is to attempt a reconciliation of the neoclassical and accelerator theories of investment de-

mand. In order for profit-maximizing behavior to generate a dependence of investment demand upon income, it is necessary to modify only one implicit assumption of the conventional neoclassical approach—namely, the assumption that the market for output is clearing. The neoclassical treatment of output as a decision variable and not as a datum implies that the neoclassical firm behaves as if the output which it plans to supply can in fact be sold. In general, this behavior is tenable only if the aggregate demand for output is at least equal to the aggregate supply of output. Alternatively, if the aggregate supply of output exceeds the aggregate demand, a representative firm would find itself unable to sell the output which it desires to supply at existing prices. Such a situation implies the treating of sales as a datum to which output and investment demand must adjust, rather than as a decision variable, even in a profit-maximizing context.

The present paper analyzes the derivation of effective investment demand functions under conditions of excess supply for current output. Within this context, a choice-theoretic approach which is consistent with the essential spirit of neoclassical theory—in particular, consistent with profit-maximizing behavior—generates demand relationships which include a dependence of investment demand upon the rate of change of output, which is the essence of the income-investment accelerator. The analysis in this paper also considers the role of expectations and the speed of capital-stock adjustment in defining the accelerator mechanism.

\* Associate professor of economics, Brown University. The National Science Foundation, Brown University, and a University of Manchester Simon Senior Research Fellowship supported this research. I am indebted to Robert Barro for suggesting the original idea for this paper. I have also benefited from the comments of Michael Parkin and Robert Coen.

The problem of the nature and significance of demand behavior under non-market-clearing conditions, considered here in relation to the theory of investment demand, is, of course, of quite general interest. Within this context, an important body of recent literature has been especially concerned with the appropriate role of income or output as arguments of demand functions in a variety of situations. In his important reappraisal of Keynesian economics, Robert Clower emphasized that this problem arises with regard to the Keynesian consumption function. Clower stressed that whereas in neoclassical utility maximization analysis, labor supply, and hence income, is a choice variable determined by the household jointly with its consumption and asset demands, in Keynesian analysis, employment and income are given data to which the household passively adjusts its consumption and asset demands. Clower explained how these divergent treatments could be reconciled by explicitly recognizing that the neoclassical model assumes the labor market to be clearing, whereas the Keynesian model assumes that there is excess supply in the labor market.

In an earlier contribution, Don Patinkin analyzed firms' demand for labor services when there is excess supply in output markets. This existence of excess supply caused the firms to regard the level of sales as a demand-determined constraint, to which they then adjusted output and labor demand. In a recent paper, Robert Barro and I integrated the Clower and Patinkin analyses to form a general model of the determination of income and employment under non-market-clearing conditions.

The approach taken in the present paper is analogous to that proposed by Clower and Patinkin. However, these earlier analyses suppressed, for simplicity, the essential intertemporal nature of the profit and utility maximization problems. Such a

simplification is, of course, out of the question in the context of analysis of investment demand. The analysis in the present paper explicitly recognizes the intertemporal nature of the profit-maximization problem, and considers, for the first time, the role of a demand-determined constraint upon sales within an explicitly intertemporal context.

In what follows, Section I reviews the conventional choice-theoretic neoclassical formulation, which apparently rules out output as an argument of the implied investment demand function. Section II then introduces excess supply in the market for current output, and explains how the conventional choice-theoretic analysis must now be modified to incorporate deficiency of demand as an additional constraint. The essential implication of this modification is the introduction of output as an argument of the implied investment demand function. Sections III and IV then go on to discuss possible forms which the accelerator relationship can take within the context of the investment demand function derived in Section II. Finally, Section V presents a brief summary of the arguments and results.

### I. A Conventional Neoclassical Model

In the neoclassical approach, the level of investment demand emerges from the profit maximizing decisions of firms which employ capital goods. In the usual example involving perfect competition, the firms are price takers in all markets. They can employ labor, issue securities, purchase capital goods, and sell their output in any quantity at prices which they take as given. However, labor is a variable factor, whereas capital is a fixed factor. More precisely, changes in the employment of capital involve a penalty cost, whereas changes in the employment of labor do not. Consequently, at every point in time (what is usually called the short run), the

firm employs labor and sells output so as to maximize current profits subject to a given stock of capital. However, over time the firm invests (that is, purchases additional capital) so as to achieve the optimal long-run (target) quantity of capital, which is that quantity dictated by current profit maximization subject to a variable stock of capital. The time path of investment chosen to achieve the target capital stock is that which maximizes the present value of the entire future stream of returns to its owners.

In order to formalize the analysis, define the following variables:

$l$  = flow of labor services

$y$  = flow of output

$K$  = stock of capital goods

$k$  = flow of capital goods, i.e., investment ( $k \equiv \dot{K}$ )

$i$  = real flow of investment expenditure, i.e., investment costed in units of output

$w$  = real wage rate

$r$  = interest rate

$V$  = present value of stream of returns to firm's owners

The superscripts  $d$  and  $s$  indicate quantities demanded and supplied, respectively. The subscript  $t$  denotes a time dependent variable. The dot indicates a time derivative. To facilitate the exposition, think of output and capital as being different goods, ignore depreciation, and assume both  $w$  and  $r$  to be constant over time.

In the analysis which follows, the firm will form expectations regarding the future values of the variables which it takes as exogenous, and on the basis of these expectations will anticipate the future time paths of its decision variables, such as  $k$ , and the state variable  $K$ . Of course, over time the firm's expectations regarding the more distant future may change. If such changes occur, the firm will form new plans, which will imply decisions different from those which the firm initially antici-

pated. Consequently, we must distinguish between the future values which the firm currently anticipates, e.g.,  $K_t$ , and the future values which actually obtain, e.g.,  $\hat{K}_t$ , which are denoted by the hat. Note that at the current moment, i.e., at time  $t=0$ , this distinction has no relevance, that is  $K_0 \equiv \hat{K}_0$ .

Using these variables, the neoclassical model sketched above may be formalized as follows:<sup>1</sup>

The firm maximizes the present value of the expected future stream of returns to its owners, which is given by

$$V = \int_0^{\infty} e^{-rt}(y_t^s - wl_t^d - i_t)dt$$

This maximization is subject to two constraints: First, the production function limits current output according to

$$y_t^s \leq F(K_t, l_t^d)$$

Assume the production function to exhibit

<sup>1</sup> See Robert Lucas' paper for a more detailed exposition. The basic reference is to Robert Eisner and Robert Strotz. A pair of possible variations on this formulation may be mentioned: One possibility would have capital users rent rather than own capital goods. In this formulation, an explicit rental rate, determined in an explicit rental market, would have to adjust at every moment to insure employment of the existing stock of capital. Given the interest rate, this rental rate would imply a demand price for capital goods. Another possibility, which is potentially a more fundamental departure, would be to regard capital as a variable factor, like labor. Formally, this treatment would involve assuming that  $I_{kk}$  (see below) equals zero. In this formulation, firms would at every moment attempt to utilize the optimal long-run stock of capital, and the demand price of capital goods at any moment would be that which made the optimal stock equal to the existing stock. Either of these variations would make for a somewhat more complex model which would focus on the determination of the demand price of capital goods rather than on the investment demand function of capital owner-users. However, neither of these variations would alter the essential arguments of this paper with regard to the role of output or its rate of change in the relevant determination. For models which incorporate both of these variations, see the articles by James Witte and Duncan Foley and Miguel Sidrauski. Note that Witte's capital users, for unspecified reasons, regard their output as exogenous, whereas the Foley-Sidrauski capital users are neoclassical.

positive and diminishing marginal product with respect to each input, and decreasing returns to scale. Second, investment expenditure depends upon the rate of investment according to<sup>2</sup>

$$i_t = I(k_t^d); \quad I(0) = 0; \\ I_k(k_t^d) > 0; \quad I_{kk}(k_t^d) > 0$$

The condition that  $I_{kk}$  is positive means that investment is subject to increasing marginal cost. This condition is consistent with the horizontal supply curve for investment goods of perfect competition if unit installation costs increase with the speed of installation of investment goods.<sup>3</sup> It is also usual to assume in this context that  $k$  cannot be negative—there is no market for used capital goods. Finally, implicit in this formulation is the assumption that expectations regarding  $w$  and the  $F$  and  $I$  functions are static. By assumption, these expectations are correct, and, consequently, will not change over time.

This maximization yields a labor demand function, an output supply function, and an investment demand function, which represents the optimal pattern of capital accumulation.<sup>4</sup> The labor demand function is

$$l_t^d = l^d(K_t, w),$$

such that

$$F_l(K_t, l_t^d) = w$$

<sup>2</sup> A function followed by a subscript indicates a partial derivative with respect to the variable in the subscript.

<sup>3</sup> In any event, this supply can be horizontal only for low rates of investment. If a firm demands a flow of investment goods which is significant relative to the aggregate flow supply of investment goods, even if it is going to maintain this rate of flow for only a short time, it must contend with a rising supply curve.

<sup>4</sup> The Euler conditions are

$$F_l(K_t, l_t^d) - w = 0$$

$$F_K(K_t, l_t^d) - rI_k(k_t^d) + \dot{k}_t^d I_{kk}(k_t^d) = 0, \text{ and } \dot{k}_t^d \geq 0$$

See Lucas for a complete spelling out of this derivation.

The output supply function is simply

$$y_t^s = F(K_t, l_t^d)$$

The investment demand function may be separated into two components. The first component is the target (long-run optimal) capital stock, denoted by  $K^*$  and given by

$$(1) \quad K^* = K^*(r, w),$$

such that

$$F_K[K^*, l^d(K^*, w)] = r \cdot I_k(0)$$

Note that the firm does not regard  $K^*$  as time dependent. The second component is the optimal time path of the capital stock, which, if we consider quadratic approximations of the  $F$  and  $I$  functions, is given by the gradual adjustment relationship

$$(2) \quad \dot{k}_t^d = \lambda(K^* - K_t)$$

where  $\lambda$  is the adjustment coefficient, and is given by

$$\lambda = -\frac{r}{2} + \left[ \frac{r^2}{4} - \frac{\frac{d^2 F}{dK^2}}{I_{kk}} \right]^{\frac{1}{2}} > 0$$

where

$$\frac{d^2 F}{dK^2} \equiv F_{KK} + F_{lK} \frac{dl^d}{dK} = F_{KK} - \frac{(F_{lK})^2}{F_{ll}}$$

According to this theory of investment, the firm plans for the capital stock to approach  $K^*$  asymptotically.<sup>5</sup> Accordingly, the firm plans for labor demand and output supply to approach asymptotically  $l^* = l^d(K^*, w)$  and  $y^* = F(K^*, l^*)$ , respectively.

<sup>5</sup> An alternative model of the adjustment process could assume that installation costs are lump sum and independent of the speed of adjustment so that  $I_{kk} = 0$ , but that  $K^*$  is taken to be stochastic rather than to be permanently fixed. For an example of such an adjustment model, see Robert Barro's article. Barro shows that, although optimal behavior in this context proximately involves an *sS* policy of all-or-nothing adjustments, average behavior can be readily approximated by a gradual adjustment relationship of the form of equation (2) but with  $\lambda$  specified somewhat differently.

The essential point to be made here is that the "neoclassical" investment demand function, obtained by substituting equation (1) into equation (2), does not have output as an argument. Given the  $F$  and  $I$  functions, investment demand depends only upon the interest rate, the real wage rate, and the existing capital stock.<sup>6</sup> In the neoclassical framework, output—both current output  $y_t^s = F(K_t, l_t^d)$  and long-run optimal output  $y^* = F(K^*, l^*)$ —are choice variables. Output is not given to the firms, but rather is determined by them in the course of their profit maximizing calculus. In other words, output is maximized out and, therefore, cannot appear as an argument of the investment demand function.<sup>7</sup>

The investment demand function is, of course, a structural relationship. In formulating its investment demand, the firm takes the real wage rate, the real cost of acquiring capital goods, and the interest rate all as given. The actual amount of investment which takes place in the neoclassical model emerges from the general equilibrium solution, in which the aggregate supply of output, demand for labor, demand for capital goods, and corresponding supply of securities by the capital-using firms are equated to the corresponding demands and supplies generated by the other sectors of the economy.

## II. Excess Supply of Output

The neoclassical approach, as represented in the preceding section, apparently

<sup>6</sup> Since  $k \equiv \dot{K}$ , equation (2) is a first-order differential equation. The solution of this equation will express  $k_t^d$  as a function of  $r$ ,  $w$ , and initial conditions.

<sup>7</sup> The absence of a causal relationship running from output to investment demand does not mean that the supply of output and the capital stock are unrelated. In particular,  $y_t^s$  does depend upon  $K_t$  and  $w$ . Moreover, given  $w$ ,  $y^*$  and  $K^*$  are uniquely related. Sometimes these relationships are confused with a dependence of  $K^*$  or  $k_t^d$  upon  $y$  or  $\dot{y}$ . Robert Coen (1969, esp. pp. 375–78), has stressed this problem and discussed some of its implications.

suggests that a choice-theoretic theory of investment demand precludes an income accelerator. The principle objective of this paper is to show this conclusion to be unwarranted. In order to generate a relationship between investment demand and income, it is necessary to modify only one assumption of the conventional neoclassical representation of perfect competition as presented above. That model assumed that the firm was a price taker and that it behaved as if its current supply of output as well as its planned future supplies of output could in fact be sold. In the present context, this last assumption is crucial. If, alternatively, the firm, taking prices as given, is unable to sell the output which it desires to supply, its investment demand will become a function of the level of output which it is able to sell.

The conventional neoclassical assumption that all output supplied can in fact be sold implies that the aggregate demand for output must at least equal the aggregate supply of output. As an alternative, suppose that, given the other variables of the system, output prices are currently sufficiently high that the current aggregate supply of output exceeds the aggregate demand.<sup>8</sup> When output is in excess supply, voluntary exchange implies that actual total sales will equal the total quantity demanded. Consequently, the representative producer will not be able to sell its supply quantity  $y_0^s$ . From the standpoint of the representative producer, the current level of sales is no longer a choice variable, but rather is a constraint, whose value is determined by the current level of demand.<sup>9</sup> Let  $y_0$  represent its actual current

<sup>8</sup> The present analysis is not concerned with rationalizing the fact that transactions take place at prices which are inconsistent with market clearing. For discussions of the theory of price adjustment, see the references to Kenneth Arrow, Barro, and Grossman (1972).

<sup>9</sup> Coen (1971) has also suggested that the failure of the output market to clear could reconcile the accelerator with profit maximization. However, in his brief

demand determined sales, where<sup>10</sup>  $y_0 < y_0^s$ .

Assume for the moment, that  $y_t = y_0$  for all  $t$ . In other words, expectations regarding the level of the demand determined constraint on sales are static. The representative firm anticipates that it will always be able to sell just output quantity  $y_0$ , and no more.<sup>11</sup> Under these conditions, maximization of the present value of expected future returns involves two calculations. First, the firm plans to produce now and in the future just output quantity  $y_0$ ,<sup>12</sup> and to do so at each date with the minimum possible amount of labor, given the stock of capital existing at that date. Second, given this strategy for employing labor, the firm chooses an optimal pattern for the accumulation of capital.

The demand for labor is then given by

$$l_t^{d'} = l^{d'}(K_t, y_0)$$

such that

$$F[K_t, l^{d'}(K_t, y_0)] = y_0$$

Denote  $l_t^{d'}$  as the "effective" demand for labor to distinguish it from the "notional"

---

discussion, pp. 143-46, Coen proposes excess demand, rather than excess supply, as a rationalization for treating the current level of sales as a constraint. This proposal implies that firms must satisfy the demand for their output at any level. Although such a restriction might be realistic in some cases, such as regulated public utilities, it is inconsistent with the assumption that exchange is voluntary. Given voluntary exchange, current sales imply a constraint upon firm behavior only in the event of excess supply.

<sup>10</sup> In principle,  $y$  need not be less than  $y_0^s$  for every firm. The apportionment of the actual sales among the firms depends upon nonprice rationing procedures. Grossman (1971) presents an explicit specification of this apportionment within a framework of voluntary exchange.

<sup>11</sup> Note that since, by assumption, capital does not depreciate,  $y_t^s$  for  $t > 0$  must be at least as large as  $y_0^s$ . Thus,  $y_0 < y_0^s$  implies  $y_t < y_t^s$  for all  $t$ .

<sup>12</sup> This analysis abstracts from inventory accumulation or decumulation. For simplicity, I assume that output always equals sales. Permitting inventory accumulation would not affect the essentials of the analyses, although it would introduce a complication analogous to the inclusion of an additional input.

demand<sup>13</sup>  $l_t^d$ . Note that, whereas the  $l^d$  function was defined by  $F_l$  equal to  $w$ , since  $y_0$  is less than  $y_t^s$ , along the  $l^{d'}$  function  $F_l$  is greater than  $w$ .

Over the long run, the firm now chooses a time path of investment to maximize

$$V = \int_0^\infty e^{-rt}(y_0 - w l_t^{d'} - i_t^{d'}) dt,$$

subject to

$$\dot{l}_t^{d'} = l^{d'}(K_t, y_0) \quad \text{and} \quad \dot{i}_t^{d'} = I(k_t^{d'})$$

Denote the variable  $k_t^{d'}$  as the effective investment demand to distinguish it from the notional demand  $k_t^d$ . This maximization yields an effective investment demand function, which also may be separated into two components.<sup>14</sup>

The first component is the effective target capital stock denoted by  $K'$ , and given by

$$(3) \quad K' = K' \left( \frac{r}{w}, y_0 \right),$$

such that

$$w \cdot l_K^{d'}(K', y_0) = -r \cdot I_k(0)$$

This specification may be expressed alternatively in terms of the production function using the relationship

$$l_K^{d'} = -F_K/F_l$$

Note that, whereas the  $K^*$  function was defined by  $F_K$  equal to  $rI_k(0)$ , since  $F_l$  now

<sup>13</sup> A similar analysis of the effective demand for labor appears in Patinkin's book. Barro and Grossman show that Patinkin's treatment of firm behavior is formally analogous to Clower's treatment of household behavior. Barro and Grossman emphasize the interaction between firm and household behavior in a general disequilibrium framework. The present paper is limited to a partial disequilibrium analysis of firm behavior, but with explicit treatment of the problem of intertemporal choice.

<sup>14</sup> The Euler condition is

$$-w l_K^{d'}(K_t, y_0) - r I_k(k_t^{d'}) + k_t^{d'} I_{kk}(k_t^{d'}) = 0,$$

and  $k_t^{d'} \geq 0$

exceeds  $w$ , along the  $K'$  function,  $F_K$  exceeds  $rI_k(0)$ . Therefore, given  $r$  and the  $F$  and  $I$  functions,  $K'$  must be smaller than  $K^*$ . This relationship reflects the fact that, given the demand imposed constraint on sales, the only motivation for expanding the capital stock is to economize on labor. The specification of the  $K'$  function reflects the optimum extent of such economizing. Note also that the firm does not regard  $K'$  as time dependent.

The second component of effective investment demand is the effective optimal time path of the capital stock. If we consider quadratic approximations of the  $I^d$  and  $I$  functions, this time path is given by the gradual adjustment relationship

$$(4) \quad \dot{k}_t^{d'} = \lambda'(K' - K_t)$$

$\lambda'$  may be denoted as the effective adjustment coefficient, and is given by

$$\lambda' = -\frac{r}{2} + \left[ \frac{r^2}{4} + \frac{w I_{KK}^{d'}}{I_{kk}} \right]^{\frac{1}{2}} > 0$$

This specification may be expressed alternatively in terms of the production function using the relationship

$$I_{KK}^{d'} = \frac{F_{IK}F_K - F_{KK}F_I}{(F_I)^2}$$

If  $K_t \geq K'$ , then  $k_t^{d'}$  will be zero. Otherwise,  $k_t^{d'}$  will be positive. In general, it is interesting to consider the relationship between  $k_t^{d'}$  and  $k_t^d$ , which depends in turn upon the relationships between  $K'$  and  $K^*$  and between  $\lambda'$  and  $\lambda$ . We have already observed that  $K'$  is unambiguously smaller than  $K^*$ . However, the relationship between  $\lambda'$  and  $\lambda$  is ambiguous. If  $F_I$  is much larger than  $w$ ,  $\lambda$  will be larger than  $\lambda'$ . Otherwise,  $\lambda'$  may be larger than  $\lambda$ . In that event,  $k_t^{d'}$  may exceed  $k_t^d$ , for small values of  $t$ . In other words, it is not necessary that a demand-imposed constraint upon output will initially reduced invest-

ment demand. However, given  $K' < K^*$ , even if  $k_t^{d'}$  initially exceeds  $k_t^d$ ,  $k_t^{d'}$  must eventually become smaller than  $k_t^d$ .

In any event, the essential point to be made here is that the effective investment demand function, obtained by substituting equation (3) into equation (4), does have current output as an argument. Current output enters through the target capital stock function as it represents the constraint upon the demand for inputs imposed by the excess supply of output. The above analysis shows that, given such a demand-imposed constraint upon output, a choice-theoretic approach yields, for a given capital stock, a direct dependence of investment demand upon the level of output. The conventional neoclassical analysis of the preceding section applies when the demand-imposed constraint upon output is not effective—that is, when the output market is not characterized by excess supply.

Like the notional investment demand function, the effective investment demand function is also a structural relationship. However, in formulating its effective investment demand, the firm takes as given the level of output, as well as the real wage rate, the real cost of acquiring capital goods, and the interest rate. Moreover, the actual amount of investment which takes place no longer emerges from a general equilibrium solution. At the least, the market-clearing condition for the output market is not satisfied. There is excess supply in that market. A plausible rationalization for this assumption is that output prices are sticky at a higher level than that required to clear the market. Also, as we have just seen, the failure of the output market to clear alters the form of the input demand functions. In particular, any disturbance which depresses output demand and creates an excess supply for output leads to depressed levels of input demands. If input prices are also down-

ward sticky, excess supplies of inputs will follow. The market interactions involved here are complex—in particular, the excess supply of inputs will feed back on the effective demand for output.<sup>15</sup> However, this paper is concerned only with a partial analysis focussed on the investment demand function. Nevertheless, we may note that in a generally depressed economy in which excess supplies characterize the markets for both outputs and inputs, effective investment demand presumably will be a constraint upon, and thereby will directly determine, the actual amount of investment.<sup>16</sup>

### III. The Income-Investment Accelerator

Equations (3) and (4) specify effective investment demand to be an increasing function of the difference between the target capital stock and the actual capital stock, and the target capital stock to be an increasing function of the level of the demand-determined constraint on output. Thus, effective investment demand depends directly upon the level of output. However, the income-investment accelerator specifies investment demand to depend upon changes in the level of output. The purpose of this section is to point out that a form of accelerator relationship, what might be called a "gradual accelerator," is implicit in equations (3) and (4).<sup>17</sup>

<sup>15</sup> We have previously analyzed these interactions in other contexts—see the articles by Barro and Grossman and by Grossman (1971).

<sup>16</sup> Of course, if investment goods and the other forms of output are identical, this last observation follows directly from the assumption of excess output supply. We may note that formulations, exemplified by the Witte and Foley-Sidrauski articles, in which capital owners do not engage in gradual adjustment, cannot readily model the determination of the actual level of investment if either output is homogeneous or excess supply exists in the capital goods market.

<sup>17</sup> Some writers refer to all gradual capital stock adjustment relationships, such as equations (2) and (4), as representing a "flexible accelerator," whether or not the target capital stock is a function of output. This terminology would seem to confuse the essential issue.

The analysis which led to equations (3) and (4) assumed that the firms had static expectations. Whatever the current level of the demand-determined constraint on output, the representative firm expected that level to prevail forever. Accordingly, it determined its current effective investment demand and planned a future time path of effective investment demand, given by  $\hat{k}_t^{d'}$ . However, if the level of the demand-determined constraint on output should actually change, the firm's expectations of the future level of this constraint would change *pari passu*, and it would alter its investment plans accordingly. Thus, if the firm's original expectations turn out to be incorrect, the actual time path of effective investment demand, denoted by  $\hat{k}_t^{d'}$ , will differ from the time path  $\hat{k}_t^{d'}$  which the firm originally anticipated. What we want to consider is how the actual time path of effective investment demand  $\hat{k}_t^{d'}$  will depend upon the actual time path of the demand-determined constraint on output  $y_t$ , given that the firm's behavior is based on static expectations. The relationship between  $\hat{k}_t^{d'}$  and  $y_t$  follows directly from equations (3) and (4) and is given by

$$(5) \quad \hat{K}_t' = K' \left( \frac{r}{w}, y_t \right)$$

and

$$(6) \quad \hat{k}_t^{d'} = \lambda' (\hat{K}_t' - \hat{K}_t)$$

Although the firm does not anticipate any change in the target capital stock over time, if  $y_t$  changes over time, the target capital stock will also change over time.

The simplest case arises when the firm's expectations are correct and, therefore, unchanging. Here  $y$  and  $\hat{K}'$  will in fact be constant over time, and the actual time paths of effective investment demand and the capital stock, i.e.,  $\hat{k}_t^{d'}$  and  $\hat{K}_t$ , will be identical to the time paths which the firm anticipates, i.e.,  $k_t^{d'}$  and  $K_t$ . Suppose fur-

ther that  $\hat{k}_t^{d'}$  is positive, so that  $\hat{K}_t$  is growing. Since  $\hat{k}_t^{d'}$  is proportionate to the difference between  $\hat{K}'$  and  $\hat{K}_t$ , as  $\hat{K}_t$  grows,  $\hat{k}_t^{d'}$  will decline. Over time, with  $\hat{y}$  (and  $r/w$ ) constant,  $\hat{k}_t^{d'}$  will asymptotically approach zero.

The essential observation here is that equation (6) represents a first-order differential equation. The above example with effective investment demand steadily declining illustrates a general implication of the solution to this differential equation. If we consider a linear approximation of the  $K'$  function—such linearity follows from our earlier quadratic approximation of the  $l^{d'}$  function—we obtain<sup>18</sup>

$$(7a) \quad \lim_{t \rightarrow \infty} \hat{k}_t^{d'} = K'_y \hat{y}_t$$

Over time, the effective absolute rate of investment demand asymptotically approaches a multiple of the absolute rate of growth of output. Alternatively, if we consider a *log*-linear approximation of the  $K'$  function—such loglinearity possesses the property that  $y=0$  implies  $K'=0$ —we obtain<sup>19</sup>

<sup>18</sup> Substitute equation (5) into equation (6) and approximate by

$$(6a) \quad \hat{k}_t^{d'} \equiv \hat{K} = \lambda'(a\hat{y}_t + b - \hat{K}_t)$$

where  $a(\equiv K'_y)$  and  $b$  are constants. Assume that  $\hat{y}_t = y_0 + ct$ , so that  $\hat{y} = c$ , where  $c$  is a constant. The solution to equation (6a) is then

$$\hat{K}_t = \left[ K_0 - a \left( y_0 + \frac{b}{a} - \frac{c}{\lambda'} \right) \right] e^{-\lambda' t} + a \left( \hat{y}_t + \frac{b}{a} - \frac{c}{\lambda'} \right)$$

which implies

$$\hat{k}_t^{d'} = -\lambda' \left[ K_0 - a \left( y_0 + \frac{b}{a} - \frac{c}{\lambda'} \right) \right] e^{-\lambda' t} + ac$$

As  $t$  becomes very large, the exponential terms vanish, and  $\hat{k}_t^{d'}$  approaches  $ac$ .

<sup>19</sup> Substitute equation (5) into equation (6) and approximate by

$$(6b) \quad \hat{k}_t^{d'} \equiv \hat{K} = \lambda'(\beta y_t^\alpha - K_t),$$

where  $\alpha(\equiv K'_y y/K)$  and  $\beta$  are constants. Assume that  $= \hat{y}_t y_0 e^{\hat{y} t}$ , so that  $(\hat{y}/y) = \hat{g}$ , where  $\hat{g}$  is a constant. The solution of equation (6b) is then

$$(7b) \quad \lim_{t \rightarrow \infty} \left( \frac{\hat{k}_t^{d'}}{\hat{K}} \right)_t = \left( K'_y \frac{y}{K} \right) \left( \frac{\hat{y}}{\hat{y}} \right)_t$$

Over time, the effective relative rate of investment demand asymptotically approaches a multiple—the coefficient here being a constant elasticity—of the relative rate of growth of output.

The asymptotic property of the accelerator relationships (7a) and (7b) should be stressed. An increase in the growth rate of output has no instantaneous effect upon effective investment demand. Rather, the effects build up only gradually over time. The speed with which the effective rate of investment demand (either absolute or relative) converges to a multiple or the growth rate of output depends directly upon the value of  $\lambda'$ .

The conventional textbook accelerator model assumes that  $\lambda'$  is very large.<sup>20</sup> This assumption implies that  $I_{kk}$  is small, that is, that investment is subject to only mildly increasing marginal costs. In this case, desired capital stock adjustment is rapid, and the asymptotic accelerator relationships (7a) and (7b) converge rapidly. The income-investment accelerator effect then builds up rapidly, and may be taken to be almost complete within a single quarterly observation period.

$$\hat{K}_t = \left( K_0 - \frac{\lambda' \beta y_0^\alpha}{\lambda' + \alpha \hat{g}} \right) e^{-\lambda' t} + \frac{\lambda' \beta}{\lambda' + \alpha \hat{g}} = \hat{y}_t^\alpha,$$

which implies

$$\hat{k}_t^{d'} = -\lambda' \left( K_0 - \frac{\lambda' \beta y_0^\alpha}{\lambda' + \alpha \hat{g}} \right) e^{-\lambda' t} + \frac{\alpha \hat{g} \lambda' \beta}{\lambda' + \alpha \hat{g}} \hat{y}_t^\alpha$$

As  $t$  becomes very large, the exponential terms vanish, and the ratio  $(\hat{k}_t^{d'}/K)_t$  approaches  $\alpha \hat{g}$ .

<sup>20</sup> The usual textbook example is expressed in terms of discrete time, that is,

$$K_t^{d'} - K_{t-\Delta} = K'_{t-\Delta} - K'_{t-2\Delta}$$

The present discussion in terms of continuous time represents the limiting case as  $\Delta$  becomes very small.

## IV. Extrapolative Expectations

The above analysis assumed throughout that expectations regarding the level of the demand-determined constraint on output were static. The representative firm always expected the current level of this constraint to prevail forever. It never anticipated changes in the level of the constraint. The preceding section pointed out that the effective investment demand function under the assumption of static expectations implies a gradual income-investment accelerator relationship. However, if the capital stock adjustment coefficient  $\lambda'$  is small, significant accelerator effects will not materialize until a large amount of time has passed.

Obviously, a number of other mechanisms for generating expectations are conceivable. This section considers one interesting possibility—that expectations are extrapolative. Changes in the level of the demand constraint generate expectations of further changes in the same direction. Consequently, expected future levels of the output are directly related to the current rate of change of output.

As a particular example of extrapolative expectations, assume that expectations are static about the relative rate of change of output. The representative firm always expects the current value  $(\dot{y}/y)_0$  to prevail forever. In this case, the firm determines its optimal pattern of capital accumulation by maximizing

$$V = \int_0^{\infty} e^{-rt}(y_t - wl_t^{d'} - i_t') dt,$$

subject to

$$l_t^{d'} = l^{d'}(K_t, y_t) \quad \text{and} \quad i_t' = I(k_t^{d'}),$$

where<sup>21</sup>

<sup>21</sup> It might be objected that

$$y_t = y_0 e^{g_0 t}$$

implies that eventually  $y_t$  must equal  $y_t^e$ , so that the

$$y_t = y_0 e^{g_0 t} \quad \text{and} \quad g_0 = (\dot{y}/y)_0$$

For the special case of  $g_0 = r$ , again considering quadratic approximations of the  $l^{d'}$  and  $I$  functions, the result of this maximization is to replace equation (4) for the planned time path of effective investment demand with<sup>22</sup>

$$(8) \quad k_t^{d''} = \lambda' \left[ K' \left( \frac{r}{w}, y_t \right) - K_t \right] + g_0 y_t K_y' \left( \frac{r}{w}, y_t \right)$$

Optimal adjustment of the capital stock now involves two components: first, gradual elimination of the gap between the current value of the effective target capital stock and the existing capital stock; second, allowance for the anticipated rate of change of the effective target capital stock.

The actual time path of effective investment demand will now be given by

$$(9) \quad \hat{k}_t^{d''} = \lambda' \left[ K' \left( \frac{r}{w}, \hat{y}_t \right) - \hat{K}_t \right] + \hat{g}_t \hat{y}_t K_y' \left( \frac{r}{w}, \hat{y}_t \right)$$

firm would no longer be constrained. An alternative assumption would be

$$y_t = y^* - (y^* - y_0)e^{-h_0 t},$$

where

$$h_0 = \left( \frac{\dot{y}}{y^* - y} \right)_0$$

On this assumption, the firm expects the demand-imposed constraint on sales to disappear asymptotically. For the special case of  $h_0 = \lambda'$ , this assumption also yields equation (8).

<sup>22</sup> The Euler condition is

$$-w l_{KK}^{d'}(K_t, y_t) - r I_k(k_t^{d''}) + \dot{k}_t^{d''} I_{kk}(k_t^{d''}) = 0, \quad \text{and} \quad k_t^{d''} \geq 0$$

If  $g_0 \neq r$ , we must add to the right side of equation (8) the following expression:

$$\frac{(g_0 - r)(g_0 + \lambda')g_0 I_{kk} y_t K_y'}{w l_{KK}^{d'} - (g_0 - r)g_0 I_{kk}}$$

The important new element here is that the rate of change of output, together with the level of output, now directly enters the effective target capital stock function, and consequently has a direct effect upon effective investment demand. Thus, an increase in the relative growth rate of output  $\hat{g}$  now has an instantaneous positive effect upon effective investment demand. In particular equation (9) implies

$$(10) \quad \frac{d\left(\frac{\hat{k}^{a''}}{\hat{K}}\right)_t}{d\hat{g}_t} = \frac{\hat{y}_t}{\hat{K}_t} \cdot K'_y\left(\frac{r}{w}, \hat{y}_t\right)$$

Extrapolative expectations imply an instantaneous accelerator which is identical to the asymptotic accelerator given by equation (7b) for the case of static expectations. In this respect, extrapolative expectations and rapid capital stock adjustment with static expectations have similar implications.

Moreover, if the firm's extrapolative expectations regarding the relative growth rate of output are correct, i.e., if  $\hat{g}_t = g_0$ , the effective relative rate of investment still behaves asymptotically according to equation (7b).<sup>23</sup> In particular, given a *log*-linear approximation of the  $K'$  function, the eventual effect of a permanent increase in  $\hat{g}_t$  is

$$(11) \quad \frac{d\left[\lim_{t \rightarrow \infty} \left(\frac{\hat{k}^{a''}}{\hat{K}}\right)_t\right]}{d\hat{g}_t} = \frac{y}{K} \cdot K'_y$$

<sup>23</sup> Approximate equation (9) by

$$(9a) \quad \hat{k}_t^{a''} \equiv \hat{K} = \lambda'(\beta \hat{y}_t^\alpha - \hat{K}_t) + \alpha \beta \hat{g} \hat{y}_t^\alpha$$

where  $\alpha$ ,  $\beta$ , and  $\hat{g} (= \hat{y}/\hat{y})$  are constants. The solution of equation (9a) is

$$\hat{K}_t = (K_0 - \beta y_0) e^{-\lambda' t} + \beta \hat{y}_t^\alpha$$

which implies

$$\hat{k}_t^{a''} = -\lambda'(K_0 - \beta y_0) e^{-\lambda' t} + \alpha \beta \hat{g} \hat{y}_t^\alpha$$

As  $t$  becomes very large, the ratio  $\hat{k}^{a''}/\hat{K}$  approaches  $\alpha \hat{g}$ .

Thus, correct extrapolative expectations do not alter the asymptotic accelerator effect, and the instantaneous and asymptotic accelerators are then identical.

## V. Summary

The purpose of this paper was to reconcile the income-investment accelerator with the neoclassical theory of the competitive firm. The essence of the accelerator is that the firm behaves as if output were a datum to which it must adjust its demands for inputs. The essential problem is to rationalize such behavior, which, in the conventional neoclassical model, appears to be inconsistent with profit maximization. My approach was to consider a phenomenon which the conventional neoclassical model implicitly rules out—the existence of excess supply in the output market. Such excess supply imposes upon the representative firm a demand-determined constraint upon sales, which effectively fixes its level of output. The above analysis considered the maximization of the present value of the future stream of returns to the firm's owners first without and then with this constraint. In both cases, optimal investment policy involves gradual adjustment of the firm's capital stock to a target value. In the former conventional neoclassical case, the firm takes only prices as given and simultaneously chooses optimal time paths for output supply and input demands, so that the one cannot be considered a function of the other. In contrast, in the latter case, when the demand-imposed constraint on sales is effective, output becomes a datum, and the target capital stock becomes an increasing function of the level of output, as well as of prices.

Within this latter context, assuming expectations about future levels of the demand-imposed constraint on output to be static, gradual capital stock adjustment implies a gradual accelerator. Investment demand asymptotically approaches a mul-

tiple of the rate of growth of output. The more rapid is the adjustment of the capital stock, the more rapidly is the full accelerator effect approached. Alternatively, this accelerator relationship becomes instantaneous if expectations are extrapolative, i.e., static about future rates of change of the demand-imposed constraint.

We thus see the income-accelerator and the conventional neoclassical theories as complementary, both being component parts of a more general choice-theoretic theory of investment demand. Profit maximizing behavior implies an income accelerator as part of the theoretical specification of investment demand when, but only when, output markets are depressed, that is to say characterized by excess supply. In contrast, the conventional neoclassical theory, which does not allow an income accelerator, is applicable when, but only when, output markets are not characterized by excess supply.

## REFERENCES

- K. J. Arrow, "Toward a Theory of Price Adjustment," in M. Abramovitz, ed., *The Allocation of Economic Resources*, Stanford 1959.
- R. J. Barro, "A Theory of Monopolistic Price Adjustment," *Rev. Econ. Stud.*, Jan. 1972, 39, 17-26.
- and H. I. Grossman, "A General Disequilibrium Model of Income and Employment," *Amer. Econ. Rev.*, Mar. 1971, 61, 82-93.
- R. W. Clower, "The Keynesian Counter-Revolution: A Theoretical Appraisal," in F. H. Hahn and F. P. R. Brechling, eds., *The Theory of Interest Rates*, London 1965.
- R. M. Coen, "Tax Policy and Investment Behavior: Comment," *Amer. Econ. Rev.*, June 1969, 59, 370-79.
- , "The Effect of Cash Flow on the Speed of Adjustment," in G. Fromm, ed., *Tax Incentives and Capital Spending*, Washington 1971.
- R. Eisner and R. Strotz, "Determinants of Business Investment," in *Impacts of Monetary Policy; Commission on Money and Credit*, Englewood Cliffs 1963.
- D. K. Foley and M. Sidrauski, "Portfolio Choice, Investment, and Growth," *Amer. Econ. Rev.*, Mar. 1970, 60, 44-63.
- H. I. Grossman, "Money, Interest, and Prices in Market Disequilibrium," *J. Polit. Econ.*, Sept./Oct. 1971, 79, 943-61.
- , "Aggregate Demand and Employment," paper read at Western Economic Association Meetings, Aug. 1972.
- R. E. Lucas, "Optimal Investment Policy and the Flexible Accelerator," *Int. Econ. Rev.*, Feb. 1967, 8, 78-85.
- D. Patinkin, *Money, Interest and Prices*, 2d ed., New York 1965, ch. 13.
- J. G. Witte, "The Microfoundations of the Social Investment Function," *J. Polit. Econ.*, Oct. 1963, 71, 441-56.

# Advertising and the Aggregate Consumption Function

By LESTER D. TAYLOR AND DANIEL WEISERBS\*

The economic effects of advertising have been a much studied and hotly debated topic for a number of years. By now, there is fairly general agreement that, *inter alia*, advertising is important as a barrier-to-entry (see Joe Bain, William Comanor and Thomas A. Wilson, Leonard Weiss) and that advertising does succeed in shifting demand for individual products (see Neil Borden, Nicholas Kaldor, Robert Dorfman and Peter Steiner, Lester Telser (1962), Kristian Palda), but there is little agreement as to the effect of advertising on aggregate consumption. John Kenneth Galbraith would have us believe that much of consumers' spending is managed from Madison Avenue,<sup>1</sup> but such a view has still to find universal acceptance.<sup>2</sup> What is surprising, however, is that no one who has been party to the rather spirited debate generated by the Tall Gentleman's thesis has seen fit to examine by econometric methods the proposition that advertising has an impact on the aggregate consumption function. To undertake this is the purpose of this paper. In a modest, yet not insignificant, way, we feel that we have made some progress. Based on an analysis of advertising expenditures in the aggregate, our results suggest that advertising does in fact tend to increase con-

sumption at the expense of saving. But as to what the causal mechanism underlying this is, we unfortunately cannot say. It may be that advertising actually succeeds in altering tastes *a la* Galbraith, but then again it may be that advertising is simply serving to bring new goods and services to the attention of consumers.

As already noted, our analysis concentrates on the effects of advertising in the aggregate, and is conducted in the framework of the state-adjustment model of Hendrik Houthakker and Lester Taylor, as applied to aggregate consumption. Following Houthakker and Taylor, two variants of the model have been employed; the first focuses on consumption, and the second on personal saving.

Section I presents a brief description of the Houthakker-Taylor (H-T) model and discusses the ways that it can be extended to accommodate advertising. This section also provides a short description of the data and methods of estimation. Sections II and III are empirical, Section II being devoted to a presentation of results and Section III to their critical evaluation. The paper is then concluded with some final observations in Section IV.

## I. The Model, Data, and Estimation

### *The Houthakker-Taylor Model*

We assume that readers are generally familiar with the logic of the H-T state-adjustment model, and therefore shall skip over details.<sup>3</sup> With time treated continuously, the structural form of the model consists of the two equations:

<sup>3</sup> Readers are referred to ch. 1 and pp. 281-93 of Houthakker and Taylor.

\* University of Michigan and Catholic University of Louvain. The research was supported by the National Science Foundation. We are indebted to William S. Comanor, Jaime Garcia dos Santos, Saul H. Hymans, Louis Philips, Richard C. Porter, Frank Stafford, and Thomas A. Wilson for comments and criticisms, and to Amy Perrone for cheerful secretarial assistance.

<sup>1</sup> See Galbraith (1967a,b).

<sup>2</sup> See, for example, the exchange between Galbraith and Robert M. Solow in *Public Interest*, and also between Robin Marris and Solow in the same publication.

$$(1) \quad q(t) = \alpha + \beta s(t) + \gamma x(t)$$

$$(2) \quad \dot{s}(t) = q(t) - \delta s(t),$$

where  $q$  and  $x$  refer to consumption expenditures and income, respectively, and  $s$  designates a state variable representing the depreciated residue (or "stocks") remaining from consumption expenditures in the past. The first equation will be referred to as the behavioral equation and the second as the depreciation equation.

The sign of  $\beta$  may be either negative or positive depending upon whether stock adjustment or habit formation predominates. Stock adjustment refers to the situation where the state variable represents a physical inventory, such as the existing stock of automobiles or kitchen durables. In this case, we expect  $\beta$  to be negative. On the other hand, if the commodity in question is a nondurable or service for which physical inventories are either of no consequence or absent altogether, the state variable must be interpreted as a psychological quantity. For the case of tobacco, for example, the more tobacco that has been consumed in the recent past, the more, everything else being constant, will be consumed in the present. In this situation, the state variable represents the accumulated force of habit, and the sign of  $\beta$  will be positive. This defines habit formation. With habit formation, the depreciation rate  $\delta$  in equation (2) measures the speed at which the habit wears off.

When the model is applied to aggregate consumption, the state variable will represent an amalgam of forces, some making for stock adjustment and others making for habit formation. On balance, those making for habit formation are the stronger, with the result that for aggregate consumption  $\beta$  is positive.<sup>4</sup>

<sup>4</sup> See pp. 281-87 of Houthakker and Taylor. Actually,  $\beta > 0$  is simply a restatement of the well-known phenomenon, studied by Tillman Brown and James Duesenberry, of inertia (or habit persistence) in the short-run consumption function.

Since the state variable is in general unobservable, it is eliminated through the use of equation (2), which connects the change in  $s$  to new expenditures and current depreciation. A translation to discrete time then yields (apart from an error term) the estimating form of the model:<sup>5</sup>

$$(3) \quad q_t = A_0 + A_1 q_{t-1} + A_2 \Delta x_t + A_3 x_{t-1},$$

where  $A_0, \dots, A_3$  are functions of  $\alpha, \beta, \gamma$ , and  $\delta$ . Once estimates of these coefficients are obtained, they can be transformed into estimates of the structural parameters via the relationships:

$$(4) \quad \alpha = \frac{A_0(2A_2 - A_3)}{A_3(A_1 + 1)}$$

$$(5) \quad \beta = \frac{2(A_1 - 1)}{A_1 + 1} + \frac{2A_3}{2A_2 - A_3}$$

$$(6) \quad \gamma = \frac{2A_2 - A_3}{A_1 + 1}$$

$$(7) \quad \delta = \frac{2A_3}{2A_2 - A_3}$$

When the model is applied to saving,  $s$  is taken to represent the accumulated stock of financial assets in which case the coefficient  $\delta$  must be set equal to zero. Letting  $y$  denote saving, the structural equations therefore become

$$(8) \quad y(t) = \alpha + \beta s(t) + \gamma x(t)$$

$$(9) \quad \dot{s}(t) = y(t),$$

and the estimating equation:

$$(10) \quad y_t = B_1 y_{t-1} + B_2 \Delta x_t$$

In this case, the structural coefficients  $\beta$  and  $\gamma$  are related to  $B_1$  and  $B_2$  through the equations:

$$(11) \quad \beta = \frac{2(B_1 - 1)}{B_1 + 1}$$

<sup>5</sup> The translation to discrete time involves: 1) integrating equations (1) and (2) over the intervals  $t$  to  $t+1$  and  $t+1$  to  $t+2$ ; 2) subtracting the former from the latter; and 3) using the mean value theorem to obtain

$$(12) \quad \gamma = \frac{2B_2}{B_1 + 1}$$

Because of the assumption that  $\delta$  is zero, the constant term,  $\alpha$ , unfortunately cannot be determined.

*Extension of the Model to  
Include Advertising*

Advertising can be brought into the model in either of two ways. The first is to include it alongside income and the state variable in the behavioral equation,<sup>6</sup>

$$(13) \quad q(t) = \alpha + \beta s(t) + \gamma x(t) + \lambda a(t),$$

where  $a(t)$  represents the flow of advertising at time  $t$ . Including advertising in the model in this way assumes 1) that the effect of advertising on consumption is direct and 2) that it operates through a flow rather than a state variable. More particularly, this procedure posits that at any point on a consumer's preference map advertising alters the marginal rate of substitution of consumption for saving, presumably in favor of greater consumption. The alteration, however, depends upon the continued flow of advertising, for if it should cease the preference map would revert to the shape it had before there was any advertising at all. Thus, with advertising in the model in this way, a permanent alteration of tastes requires that the flow of advertising be sustained.<sup>7</sup>

The second way to include advertising in the model is via the depreciation rela-

tionship. In this case, the assumption would be that advertising affects consumption indirectly through the state variable, rather than directly through its flow as in equation (13). There are two ways that this assumption might be made operational.

One way would be to assume that advertising leads to the creation of a goodwill stock which depreciates over time. This is the approach which is typically followed in analyzing advertising from the standpoint of the firm,<sup>8</sup> and in the present context can be represented as follows:

$$(14) \quad \dot{s}(t) = \kappa_1 q(t) + \kappa_2 a(t) - \delta s(t)$$

This equation says that additions to the state variable now consist of two components—one arising from purchase of the good itself, and the other from advertising. Once part of the state variable, however, these two components become indistinguishable in the sense that they depreciate at the same rate and exert a common influence on consumption. However, since our maintained hypothesis is that advertising leads to increased consumption, we must take care to distinguish two cases, depending upon whether the specific good being advertised is subject to habit formation or to stock adjustment. If it is subject to habit formation (i.e.,  $\beta > 0$ ), then since, by definition, expenditure and the state variable *for that good* vary directly,  $\kappa_2$  must be positive. On the other hand, if the good being advertised is characterized by stock adjustment (i.e.,  $\beta < 0$ ) then since, again by definition, expenditure and the state variable *for that good* vary inversely,  $\kappa_2$  must be negative.<sup>9</sup> Thus, it is evident that the effect of advertising on the state variable will be asymmetrical, depending upon the type of

an approximation for the terms involving  $s$ . For the details, see pp. 13–17 of Houthakker and Taylor.

<sup>6</sup> For the moment, we shall confine the discussion to the consumption model.

<sup>7</sup> Unfortunately, even if the data were strongly to support this formulation of the model (as we shall see is in fact the case), we should not be able to conclude that advertising, though it affects consumption, does so through the alteration of tastes. For it may be that advertising simply serves to bring new goods to the attention of consumers, and it is the availability of these new goods that leads to the increased consumption.

<sup>8</sup> See Marc Nerlove and Kenneth Arrow. A monograph on advertising, now nearing completion, by Comanor and Wilson also employs this procedure.

<sup>9</sup> The coefficient  $\kappa_1$  will, of course, be positive for both types of goods. Indeed, we can without loss of generality simply take  $\kappa_1$  to be 1.

underlying state-adjustment behavior involved.

But here we encounter difficulty, for in order to take into account the asymmetry of the sign of  $\kappa_2$ , we require two depreciation equations:

$$(15) \quad \begin{aligned} \dot{s}_1(t) &= \kappa_{11}q_1(t) + \kappa_{12}a_1(t) - \delta_1s_1(t) \\ \dot{s}_2(t) &= \kappa_{21}q_2(t) + \kappa_{22}a_2(t) - \delta_2s_2(t), \end{aligned}$$

where the subscripts 1 and 2 refer to goods subject to habit formation and stock adjustment, respectively, and a corresponding extension of the behavioral equation:

$$(16) \quad q(t) = \alpha + \beta_1s_1(t) + \beta_2s_2(t) + \gamma x(t)$$

However, since only one of the unobservable state variables can be eliminated from the estimating equation, this particular variant of the model cannot be applied to aggregate consumption. (It can, however, be applied to individual commodities.)

The alternative to assuming the creation of a goodwill stock is to postulate that advertising operates on the state variable through the depreciation rate, in which case, we would have

$$(17) \quad \dot{s}(t) = q(t) - \delta(a(t))s(t)$$

The simplest mechanism would be to assume  $\delta(a(t))$  to be linear, i.e.,

$$(18) \quad \delta(a(t)) = \delta_0 + \delta_1a(t)$$

Once again, if the hypothesis that advertising leads to increased consumption is to be maintained, we must distinguish two cases. For goods subject to habit formation, advertisers will clearly want to minimize the depreciation rate, since it measures the rate at which the habit wears off. On the other hand, for goods such as autos or refrigerators for which the state variable represents physical inventories, the goal of advertising will be to increase the rate of depreciation, or more accurately, the rate of obsolescence. Thus, we once again have an asymmetry, for  $\delta_1$  must be negative with habit formation, but positive with

stock adjustment. As a consequence, for the same reasons as with the goodwill stock, this approach also is a loss as far as aggregate consumption is concerned.<sup>10</sup>

We are thus led to the conclusion that the only feasible way of incorporating advertising into the consumption model is to include it alongside income and the state variable in the behavioral equation. With the saving model, however, the situation seems somewhat happier. With saving, the state variable is assumed to represent the accumulated stock of financial assets, and since the behavior of saving is homogeneous with respect to the components of this stock, there does not appear to be a problem of asymmetry. Consequently, not only can advertising be included in the behavioral equation, but a model can also be estimated in which advertising enters—via the creation of a goodwill stock—into the depreciation relationship as well.<sup>11</sup>

Let us now turn to the estimating equations for the models that have been estimated. These are four in number and are as follows:

#### Model 1

Advertising in the behavioral equation with consumption as the dependent variable, viz:

$$(13) \quad q(t) = \alpha + \beta s(t) + \gamma x(t) + \lambda a(t)$$

$$(2) \quad \dot{s}(t) = q(t) - \delta s(t)$$

The estimating equation for this model is:

$$(19) \quad \begin{aligned} q_t = & A_0 + A_1q_{t-1} + A_2\Delta x_t + A_3x_{t-1} \\ & + A_4\Delta a_t + A_5a_{t-1}, \end{aligned}$$

<sup>10</sup> This variant suffers a second practical problem in that its estimating equation is highly non-linear. Consequently, its applicability even to individual commodities is not straightforward.

<sup>11</sup> However, as we shall see in Section II, the empirical results for this model are very negative, suggesting that asymmetry may still be a problem. See fn. 21 below.

where  $\alpha$ ,  $\beta$ ,  $\gamma$ , and  $\delta$  are connected to  $A_0$ ,  $\dots$ ,  $A_3$  as in (4)–(7) and  $\lambda$  is given by

$$(20) \quad \lambda = \frac{2A_4 - A_5}{A_1 + 1}$$

However,  $\delta$  is now overidentified and (19) must therefore be estimated under the restriction:<sup>12</sup>

$$(21) \quad A_2A_5 - A_3A_4 = 0$$

### Model 2

Advertising in the behavioral equation with saving as the dependent variable, viz:

$$(22) \quad y(t) = \alpha + \beta s(t) + \gamma x(t) + \lambda a(t)$$

$$(9) \quad \dot{s}(t) = y(t)$$

The estimating equation for this model is

$$(23) \quad y_t = B_1y_{t-1} + B_2\Delta x_t + B_3\Delta a_t$$

The coefficients  $\beta$  and  $\gamma$  are still given by (11) and (12), while  $\lambda$  is connected to  $B_3$  and  $B_1$  by

$$(24) \quad \lambda = \frac{2B_3}{B_1 + 1}$$

### Model 3

Advertising in the depreciation equation with saving as the dependent variable, viz:

$$(8) \quad y(t) = \alpha + \beta s(t) + \gamma x(t)$$

$$(25) \quad \dot{s}(t) = y(t) + \kappa a(t)$$

The estimating equation for this model has the form:

$$(26) \quad y_t = B_1y_{t-1} + B_2\Delta x_t + B_3(a_t + a_{t-1})$$

As before,  $\beta$  and  $\gamma$  are given by (11) and (12), while  $\kappa$  is given by

$$(27) \quad \kappa = \frac{2B_3}{B_1 - 1}$$

<sup>12</sup> Imposition of this identifying restriction implies that the path of dynamic adjustment of consumption for a change in advertising is the same as for a change in income.

### Model 4

Advertising in both the behavioral equation and the depreciation equation with saving as the dependent variable, viz:

$$(13) \quad y(t) = \alpha + \beta s(t) + \gamma x(t) + \lambda a(t)$$

$$(25) \quad \dot{s}(t) = y(t) + \kappa a(t)$$

The estimating equation for this model is:

$$(28) \quad y_t = B_1y_{t-1} + B_2\Delta x_t + B_3a_t + B_4a_{t-1}$$

Once again, (11) and (12) remain valid for  $\beta$  and  $\gamma$ ;  $\kappa$  and  $\lambda$ , however, are now given by

$$(29) \quad \kappa = \frac{B_3 + B_4}{B_1 + 1}$$

$$(30) \quad \lambda = \frac{B_3 - B_4}{B_1 + 1}$$

### Data and Estimation

The data used for advertising are taken from the *Statistical Abstract of the United States* and refer to the total of all advertising expenditure in the economy. The consumption, saving, and disposable income data are taken from the National Income Accounts,<sup>13</sup> and are expressed in 1958 dollars per capita. Since the advertising data are available only annually, the models estimated are annual and cover the period 1929–68 (with World War II years 1942–45 being omitted for obvious reasons). The identifying condition for  $\delta$  expressed in equation (21) requires non-linear estimation of equation (19), and this has proceeded via the algorithm suggested by Donald Marquardt. Apart from this, since there are no apparent problems with autocorrelation, ordinary least squares

<sup>13</sup> A complete listing of the data used and their sources will be provided by the authors upon request. We should mention that our definition of personal disposable income differs from that of the Department of Commerce in that we do not include personal transfer payments.

has been used in estimation throughout.<sup>14</sup> Approximate standard errors for the structural coefficients have been derived from the variances and covariances of the coefficients in the estimating equation using the procedure outlined in chapter 2 of Houthakker and Taylor.

Before turning to the empirical results, there are a few points that ought to be mentioned with regard to the advertising data:

1) Given that our concern is with the effect of advertising on consumption and saving, that part of advertising expenditure that is not directed toward consumers should in principle be excluded. To do this, however, would require using data from the *Statistics of Income* of the Internal Revenue Service, and these are not available prior to 1947. The loss of the prewar years, in our opinion, is too high a price to pay for the sanitization.

2) In principle, expenditures for advertising whose primary purpose is to alter tastes should be distinguished from those expenditures whose purpose is simply to inform. But with the data now available, this task is hopeless even to attempt.

3) Moreover, even if coverage were not a problem, expenditure data are hardly the ideal measure of advertising activity; the number of advertising messages received is the appropriate quantity to employ. However, once again we must be realistic as to what the data can provide.<sup>15</sup>

<sup>14</sup> In principle, generalized least squares should be used in estimation since the error terms for the estimating equations of the models are two-period moving averages involving  $\delta$  as a parameter. However, in view of the lack of autocorrelation in the *OLS* residuals (as evidenced by the new test for autocorrelation for models with lagged dependent variables recently suggested by James Durbin), the negative one-period moving-average autocorrelation introduced by the elimination of  $s$  appears approximately to be offsetting positive autocorrelation in the original error term of the structural equation.

<sup>15</sup> For a discussion of the problems, see Telser (1962).

4) Next, there is the question of correction for price changes. In the absence of a more appropriate deflator available over the entire period of the sample, our procedure has been to deflate the advertising data with the implicit deflator for *GNP*. On the other hand, in view of the argument made by Telser (1962) that expenditure data in current prices are a better measure of the number of advertising messages received,<sup>16</sup> we have also estimated models with the data undeflated.

5) Finally, there is the question of deflation as regards the population. Since (unlike income) advertising is not specific to the individual, but common to large groups of consumers, the advertising data should probably be analyzed as an aggregate. However, since it makes for easier interpretation of coefficients, most of the empirical work has been with advertising expenditures expressed in per capita terms.

## II. Empirical Results

The empirical results are tabulated in Tables 1 and 2. Table 1 gives the coefficients, standard errors, etc. for the estimating equations, while Table 2 does the same for the structural equations. In these tables, the consumption equations are designated by  $q$ , while the saving equations are designated by  $y$ . As noted in footnote 14, we have employed the new test for autocorrelation for models with lagged dependent variables as predictors, recently suggested by Durbin. In all cases, the null hypothesis of no autocorrelation could not be rejected. We shall proceed with a short discussion of each of the models in turn.

### Model 1

The equations corresponding to Model 1 are (q.1) and (q.2). With (q.1), we have a

<sup>16</sup> His reasoning is that since price changes have generally been upward, these tend to offset the improved efficiency of the advertising media.

TABLE 1—COEFFICIENTS FOR ESTIMATING EQUATIONS

Equation	Dependent Variable	Constant	$q_{t-1}$	$y_{t-1}$	$\Delta x_t$	$x_{t-1}$	$\Delta a_t$	$a_t$	$a_{t-1}$	$a_t + a_{t-1}$	$\bar{R}^2$	$S_e$
Advertising in 1958 dollars per capita												
(q.1)	$q_t$	-3.89 (14.75)	.9180 (.0968)		.4762	.0654 (.0639)	4.6900 (1.1700)		.644 (.687)		.999	15.67
(q.2)	$q_t$	1.67 (19.29)	.6920 (.1757)		.6414 (.0557)	.2940 (.1564)					.999	19.12
(y.1)	$y_t$			.8889 (.0284)	.5057 (.0466)		-3.8863 (.8928)				.903	15.41
(y.2)	$y_t$			.8836 (.0355)	.3896 (.0477)						.857	19.25
(y.3)	$y_t$			.7114 (.1063)	.3624 (.0489)					.1812 (.1058)	.857	18.70
(y.4)	$y_t$			.9190 (.1045)	.5155 (.0576)			-4.0900 (1.1332)	4.0273 (1.0212)		.901	15.64
Advertising in current dollars per capita												
(y.5)	$y_t$			.9594 (.0357)	.4935 (.0482)		-4.8270 (1.2566)				.894	16.10
Advertising in 1958 dollars (aggregate)												
(y.6)	$y_t$			.9379 (.0298)	.5156 (.0455)		-.0297 (.0063)				.909	14.90
Advertising in current dollars (aggregate)												
(y.7)	$y_t$			1.0163 (.0439)	.5011 (.0479)		-.0323 (.0080)				.898	15.82

Note: Standard errors are in parentheses.

first definite suggestion of something of importance, for not only is the sign of advertising positive, but one of its coefficients in the estimating equation is four times its standard error.<sup>17</sup> Moreover, and more importantly, it is evident from the increase in the standard error of the estimate when it is excluded as a predictor (equation (q.2)) that advertising, despite the fact that it and disposable income are highly

correlated,<sup>18</sup> is not simply "freeloading" on income and lagged consumption.

Despite these positive features, however, the overall quality of this model leaves much to be desired. The coefficients of  $x_{t-1}$  and  $a_{t-1}$  are both insignificant, and this is true also of the constant, which means, as a consequence, that it cannot be assumed that  $\delta$ , the depreciation rate in the structural model, is different from zero.<sup>19</sup> Thus, the data quite clearly point

<sup>17</sup> Unfortunately, the program used in estimating (q.1) does not give a standard error for the coefficient of  $\Delta x_t$ . Nor does it provide the coefficients' variance-covariance matrix. Therefore it has not been possible to calculate approximate standard errors for the structural coefficients.

<sup>18</sup> The correlation between advertising and income in levels is .98. In first differences, the correlation drops to .64.

<sup>19</sup> That this follows is evident from equations (1.39) and (1.60) of Houthakker and Taylor.

TABLE 2—COEFFICIENTS FOR STRUCTURAL EQUATIONS

Equation	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\gamma}'$	$\hat{\lambda}$	$\hat{\lambda}'$	$\hat{\kappa}$	$\hat{\delta}$
(q.1)	-27.53	.062	.463	.800	4.555	7.855		.148
(q.2)	3.32 (37.18)	.231 (.153)	.584 (.061)	.955 (.060)				.595 (.395)
(y.1)		-.118 (.031)	.535 (.052)		-4.115 (.950)			0
(y.2)		-.124 (.040)	.414 (.054)					0
(y.3)		-.337 (.145)	.424 (.058)				-1.255 (.348)	0
(y.4)		-.084 (.113)	.537 (.052)		-4.230 (1.015)		.755 (3.477)	0
(y.5)		-.041 (.037)	.504 (.049)		-4.927 (1.335)			0
(y.6)		-.064 (.032)	.532 (.048)		-.031 (.007)			0
(y.7)		.016 (.043)	.497 (.046)		-.032 (.008)			0

*Note:* Approximate standard errors are in parentheses. Because of the non-linear estimation, these are not available for (q.1). The coefficients  $\gamma'$  and  $\lambda'$  refer to long-run (steady-state) equilibrium, and differ from their short-run counterparts,  $\gamma$  and  $\lambda$ , by the factor  $\hat{\delta}/(\hat{\delta}-\hat{\beta})$ .

to the model with saving as the dependent variable.<sup>20</sup>

### Model 2

The equations for Model 2 are (y.1), (y.2), (y.5), (y.6), and (y.7). Quite clearly, this model leads to improved results. Focussing for the moment on (y.1), which has advertising expressed in 1958 dollars per capita, we see that all coefficients in the model—in the structural equation as well as in the estimating equation—are

multiples of their standard errors and that in keeping with a priori expectations, both  $\hat{\beta}$  and  $\hat{\lambda}$  are negative. (Reassuringly,  $\hat{\gamma}$ , the short-run marginal propensity to save also has the proper sign.) While the  $\bar{R}^2$ s of (q.1) and (y.1) are not conceptually comparable, the standard errors of the estimate are, and this being the case, we see that (y.1) has the slight edge.

From a comparison of equations (y.1) and (y.2), it is once again evident—if it were not already apparent from the  $t$ -ratio of nearly four for the coefficient of  $\Delta a_t$ —that advertising has an independent contribution of its own. While the reduction in the standard error of the estimate in going from (y.2) to (y.1) in essence is simply a repeat performance of (q.2) and

<sup>20</sup> It will be noted from (q.2) that upon the exclusion of advertising, the  $t$ -ratio for  $x_{t-1}$  still falls short of two, while the plight of the constant term actually worsens. It was this result which originally led H-T to consider treating saving as the acquisition of a nondepreciating asset.

(q.1), the effect is more dramatic with (y.2) and (y.1) since it now also shows up in the  $\bar{R}^2$ .

Turning next to the alternative definitions of the advertising variable, it is seen from a comparison of equations (y.1), (y.5), (y.6), and (y.7) that the best results, not only in  $\bar{R}^2$  but also in terms of the largest  $t$ -ratio for the coefficient of  $\Delta a_t$ , are obtained when the advertising data are corrected for price changes but left as an aggregate (equation (y.6)). However, advertising defined in this way gives results only marginally better than when advertising is expressed in constant dollars per capita (equation (y.1)), and the latter in turn is only marginally better than the equation in which the data are left in current dollars but expressed per capita (equation (y.5)). The only combination of deflation procedures leading to unacceptable results is the case where the data are left as a current dollar aggregate, for as is seen in equation (y.7), this definition of advertising leads to an estimate of  $B_1$  which is greater than one.

#### Model 3

Model 3, it will be recalled, allows for the impact on consumption (and saving) to operate via the depreciation relationship rather than through the behavioral equation as in Models 1 and 2. As can be seen from the equation (y.3), the results for this model are very negative. Not only is the coefficient of the advertising term in the estimating equation insignificant by conventional standards, but what is even worse the sign for  $\kappa$  is wrong. A negative value for  $\kappa$  implies that advertising leads to a reduction in the accumulated stock of savings, which, given that  $\beta$  is negative, means that increased advertising leads to increased saving, and therefore to reduced consumption. This is, of course, contrary to what is expected a priori.

#### Model 4

The equation corresponding to Model 4 is (y.4). In addition to being a valid model in its own right, this model is of interest because it can also be interpreted as a direct test of Model 2 vis-à-vis Model 3. That this is the case is evident from equations (23), (26), and (28): for  $B_3$  equal to  $-B_4$ , we have Model 2, while for  $B_3$  equal to  $B_4$ , we have Model 3.

From (y.4), we see that the results support neither Model 4, nor Model 3—since  $\kappa$  in Table 2, though now positive (unlike in (y.3)) is only a fraction of its approximate standard error—but unequivocally support Model 2, for equation (y.4) is virtually identical with (y.1). This result is of some importance, for it strongly supports the hypothesis that advertising affects consumption directly rather than through the intermediary of a state variable and it suggests, also quite strongly, that advertisings' effect is temporary in duration.<sup>21</sup>

#### Interpretation

At this point, it will be useful to put the foregoing in perspective. There is little question but that the results in Tables 1 and 2 are very favorable to the view that advertising has a positive impact on con-

<sup>21</sup> As noted in fn. 11, the poor performance of Model 3 may be due to the asymmetry problem that was discussed in Section II in connection with the inclusion of advertising in the depreciation relationship in the consumption model. With Model 3, the implicit assumption is that advertising affects the state variable for saving, but this is obviously pretty far-fetched, for the theme of advertising is usually "buy" rather than "don't save". This distinction is unimportant when the assumption is that the effect of advertising is direct rather than through the state variable since consumption and saving are connected via the budget constraint. However, because the state variable for consumption is conceptually different from the state variable for saving, the two state variables do not, unlike consumption and saving, satisfy any budget constraint. Consequently, to write the depreciation relationship as in (25) is clearly a misspecification, and probably a serious one.

sumption. However, we should not conclude from this that Galbraith is right, that a substantial component of consumers' spending is in fact directed by Madison Avenue. Such a conclusion is unwarranted for at least three reasons.

First, the result may be spurious in that our advertising variable may simply be standing as a proxy for some other factor. Second, it may be that the causality runs from final sales to advertising, rather than the other way around. Finally, even if we put the first two contingencies to the side—as we shall until the next section—the empirical evidence does not necessarily support the hypothesis that advertising persuades. Not all advertising is persuasive in purpose. Some is merely meant to inform, and this may be the aspect of advertising that the results are reflecting.<sup>22</sup>

Let us now turn to the quantitative importance of advertising in raising the level of consumption. Focussing first on the model with consumption as the dependent variable, we see from (q.1) in Table 2 that a one dollar per capita increase in advertising expenditures is estimated to lead to an increase in per capita consumption of about \$4.55 in the short run and \$7.85 in the long run.<sup>23</sup> While at first glance, these figures might seem large, it must be kept in mind that an ultimate increase in sales of nearly \$8 for each additional dollar of advertising says nothing about whether the advertising expenditure is profitable. Profitability will depend, among other things, upon the profit margin on sales, the length of the payout period over which the advertising expenditure (viewed as an investment to the firm) is calculated, and the rate at which the goodwill stock created by the advertising

depreciates. However, to delve into these factors, while interesting, would lead us far afield.

With regard to Model 2, we see from (y.1) in Table 2 that the short-run effect of a dollar increase in advertising is to reduce saving by about \$4.10. Because of the assumption  $\delta=0$ , however, there is no long-run response corresponding to steady-state equilibrium. Consequently, we must consider instead the equilibrium saving-income ratio which arises from steady (exponential) growth in income and advertising. To obtain the expression for this quantity, we begin by assuming that disposable income and advertising each grow at the constant (exponential) rate  $\rho$ . It then follows from the linearity of (22) and (9) that saving also will grow at this rate, i.e.,

$$(31) \quad \frac{\dot{y}}{y} = \rho$$

We now differentiate (22) with respect to time, obtaining:

$$(32) \quad \dot{y} = \beta \dot{s} + \gamma \dot{x} + \lambda \dot{a} \\ = \beta y + \gamma \hat{x} + \lambda \hat{a} \quad [\text{from (9)}],$$

which in view of (31) and our assumption about the growth of income and advertising can be written as

$$(33) \quad \rho \hat{y} = \beta \hat{y} + \gamma \hat{x} + \lambda \hat{a}$$

The "hats" are to denote that we are dealing with "golden-age" growth. Hence

$$(34) \quad \hat{y} = \frac{\rho}{\rho - \beta} (\gamma \hat{x} + \lambda \hat{a})$$

Finally, division by  $\hat{x}$  yields

$$(35) \quad \frac{\hat{y}}{\hat{x}} = \frac{\rho}{\rho - \beta} (\gamma + \lambda \sigma),$$

where  $\sigma$  denotes the (constant) advertising-income ratio. For purposes of calculat-

<sup>22</sup> See fn. 7 above.

<sup>23</sup> The short run is defined by the condition that stocks are held constant, while the long run corresponds to steady-state equilibrium.

ing the long-run effect of advertising on saving, we also require the golden-age saving-income ratio on the assumption that the growth rate for advertising is zero. From (32) and (33), this is seen to be

$$(36) \quad \frac{\hat{y}}{\hat{x}} = \frac{\gamma\rho}{\rho - \beta}$$

Using  $\hat{\beta}$ ,  $\hat{\gamma}$ , and  $\hat{\lambda}$  from equation (y.1) in Table 2 and assuming values of .025 for  $\rho$  and .032—the mean advertising-disposable income ratio observed during 1961–68—for  $\sigma$ , the value for  $\hat{y}/\hat{x}$  yielded by (35) is .070. This compares with a value of .093, obtained from (36), when the growth rate for advertising is assumed to be zero. Thus, the difference on the saving rate between steady exponential growth of 2.5 percent per annum in advertising and no growth is indicated to be 2.3 percentage points, hardly an insignificant difference.

### III. A Critical Evaluation of the Results<sup>24</sup>

In this section, we subject the results of the preceding section, which thus far we have taken at face value, to a detailed statistical critique. In particular, we shall examine the following questions:

1) As just noted, it may be that the relationship between advertising and consumption is one of simultaneity rather than unidirectional from advertising to consumption as has been postulated. If this is the case, then the coefficients presented in Table 1 are subject to simultaneous-equations bias. To test for this, we have specified an equation for advertising with consumption as an argument, and then estimated equation (y.1) by two-stage least squares.

2) The historical period in the analysis has been 1929–68. There is some legitimate concern that to include the prewar

years in the sample is improper because of possible shifts in the underlying structure. To test whether this might be the case, we have undertaken an analysis of covariance (Chow test)<sup>25</sup> with the period broken at World War II.

3) Finally, an examination of the advertising-disposable income ratio over the historical period indicates that, while there is no apparent trend, advertising is cyclically the more energetic. Consequently, it may be that advertising is merely acting as a proxy for some cyclical effect—the ebb and flow of transfer payments, for example. To test for this, we have estimated Model 2 with the unemployment rate included as a predictor along with advertising and income.

#### *Two-Stage Least Squares Estimation of Model 2*

In testing for possible simultaneous-equations bias, we have reestimated Model 2 using two-stage least squares (TSLS). In doing this, we have postulated the following model for advertising:

$$(37) \quad a_t = b_0 + b_1 a_{t-1} + b_2 \Delta q_t + b_3 q_{t-1} + \epsilon_t$$

Since we are not required actually to estimate this model, but only the reduced-form equation for  $a_t$  corresponding to it and equation (23), we need not linger over a rationale, and shall proceed instead directly to the estimated reduced form. This is as follows:

$$(38) \quad \hat{a}_t = -5.60 + .5197a_{t-1} \\ (2.50) \quad (.0885) \\ - .0075x_{t-1} + .0285q_{t-1} \\ (.0198) \quad (.0228)$$

$$\bar{R}^2 = .939 \quad S_e = 2.51$$

With this estimate of  $\hat{a}_t$ , the second stage in TSLS then yields for the estimating

<sup>24</sup> This section owes much to critical questions raised by Saul H. Hymans.

<sup>25</sup> For a description of this test, see Gregory Chow.

equation and structural coefficients for Model 2:

$$(39) \hat{y}_t = .8983y_{t-1} + .4256\Delta x_t - 2.5738\Delta a_t$$

(.0349)            (.0492)            (1.3127)

$$\bar{R}^2 = .726 \quad S_e = 18.45$$

$$\hat{\beta} = -.107 \quad \hat{\gamma} = .448 \quad \hat{\lambda} = -2.712$$

(.039)            (.054)            (1.393)

In comparing this equation with (y.1) in Table 1, we see that the standard error of the estimate is about \$3 higher than in (y.1), and that the coefficient (in absolute value) of advertising has been reduced by about one-third and is now only about twice its estimated asymptotic standard error. However, in view of the fact that standard errors for *TSLS* for finite samples do not always exist,<sup>26</sup> we should not attach much importance, one way or another, to this apparent drop in significance. In terms of the steady-growth saving-income ratios, formulas (37) and (38) now yield values of .085 and .068, respectively.

In view of these results, simultaneity between consumption (or, more particularly, final sales) and advertising is something that quite clearly cannot be ruled out. However, despite the fact that the quantitative impact of advertising has decreased, it has hardly disappeared altogether. A 1.7 percentage point difference in the golden-age saving rate is still not something to be dismissed out of hand.

#### *A Test for Coefficient Stability Between Pre- and Postwar Years*

An analysis of covariance test for the stability of the coefficients between the prewar years and those of the postwar is recorded in Table 3. The equations (in addition, of course, to (y.1)) are:

1930-41

$$(40) \hat{y}_t = .8986y_{t-1} + .5326\Delta x_t - 3.7303\Delta a_t$$

(.1094)            (.0638)            (1.2977)

$$\bar{R}^2 = .969 \quad S_e = 14.05$$

<sup>26</sup> See Robert Basman (1963a,b).

TABLE 3—TEST FOR STABILITY OF COEFFICIENTS BETWEEN PRE- AND POSTWAR YEARS

Equation	Residual Sum of Squares (RSS)	Degrees of Freedom	Mean Square
(y.1) 1930-68	7362.59	31	237.50
(38) 1930-41	1777.87	9	197.54
(39) 1947-68	5098.02	19	268.32
Sum of RSS of "within" regres- sions	6875.89	28	245.57
Reduction in RSS due to different regressions	486.70	3	162.23
$F = \frac{245.57}{162.23} = 1.51$			
$F_{.05}(28, 3) \cong 8.65$			

1947-68

$$(41) \hat{y}_t = .9221y_{t-1} + .4511\Delta x_t - 5.1137\Delta a_t$$

(.0408)            (.0732)            (1.7353)

$$\bar{R}^2 = .853 \quad S_e = 16.38$$

To reject the hypothesis that the two periods are homogeneous with regard to structure requires (at the .05 level of significance with 28 and 3 degrees of freedom) an *F*-ratio of at least 8.65. Consequently, since the observed *F* is only a fraction of this value, we cannot reject the hypothesis that the two periods have a common structure.

#### *Advertising as a Proxy for Some Cyclical Factor?*

The final contingency that we have examined is whether advertising might simply be a stand-in for some cyclical factor. To test for this, we have selected the aggregate unemployment rate as a representative cyclical variable and included it as a predictor in addition to disposable income and advertising, obtaining the following equation:

$$(42) \quad \hat{y}_t = .8958y_{t-1} + .4910\Delta x_t \\ (.0355) \quad (.0648)$$

$$- 4.0547\Delta a_t - .6208\Delta u_t \\ (1.0398) \quad (1.8817)$$

$$\bar{R}^2 = .912 \quad S_e = 15.64$$

As can be seen, the coefficient of  $\Delta u$  has the wrong sign and is only a fraction of its standard error. The coefficient of  $\Delta a_t$ , on the other hand, is scarcely affected. Consequently, it seems safe to conclude that, at a minimum, advertising is not a proxy for aggregate unemployment.

#### IV. Concluding Remarks

As was noted in the introduction, we feel that we have made modest, though nevertheless definite, progress toward resolution of the question of whether advertising affects the aggregate consumption function. The results with Model 2 quite clearly indicate that it does. However, we certainly do not wish to leave the impression that the results reported here have settled the issue, for this is quite clearly not the case.

First, we have analyzed only aggregate data and it could be that our results reflect nothing more than errors of aggregation. Second, the results of the preceding section suggest that the relationship between advertising and consumption is not unidirectional, but simultaneous. In future work, this should be dealt with more adequately than has been the case here. Next, we should note that while the preceding section shows that advertising is not acting as a proxy for the unemployment rate, it nevertheless could be standing as a proxy for something else, a "new goods" effect perhaps. And finally, we should mention once again that our analysis has been pursued entirely within the framework of the Houthakker-Taylor state-adjustment model. Another model might lead to differ-

ent conclusions, although our opinion is that this would not be the case.

#### REFERENCES

- J. Bain, *Barriers to New Competition*, Cambridge, Mass. 1956.
- R. L. Basmann, "A Note on the Exact Finite Sample Frequency Functions of Generalized Classical Linear Estimators in a Leading Three-Equation Case," *J. Amer. Statist. Ass.*, Mar. 1963, 58, 161-71.
- , "Remarks Concerning the Application of Exact Finite Sample Distribution Functions of GCL Estimators in Econometric Statistical Inference," *J. Amer. Statist. Ass.*, Dec. 1963, 58, 943-76.
- N. H. Borden, *The Economic Effects of Advertising*, Homewood 1942.
- T. M. Brown, "Habit, Persistence and Lags in Consumer Behavior," *Econometrica*, July 1952, 20, 355-71.
- G. C. Chow, "Tests of Equality Between Sets of Coefficients in Two Linear Regressions," *Econometrica*, July 1960, 28, 591-602.
- W. S. Comanor and T. A. Wilson, "Advertising, Market Structure, and Performance," *Rev. Econ. Statist.*, Nov. 1967, 49, 423-40.
- and ———, *Advertising and Market Power*, Cambridge, Mass., forthcoming.
- R. Dorfman and P. O. Steiner, "Optimal Advertising and Optimal Quality," *Amer. Econ. Rev.*, Dec. 1954, 44, 826-36.
- J. S. Duesenberry, *Income, Saving, and the Theory of Consumer Behavior*, Cambridge, Mass. 1949.
- J. Durbin, "Testing for Serial Correlation in Least-Squares Regression When Some of the Regressors are Lagged Dependent Variables," *Econometrica*, May 1970, 38, 410-21.
- H. S. Houthakker and L. D. Taylor, *Consumer Demand in the United States*, 2d ed., Cambridge, Mass. 1970.
- J. K. Galbraith, (1967a) *The Affluent Society*, 2d ed., Boston 1967.
- , (1967b) *The New Industrial State*, Boston 1967.
- , (1967c) "Review of a Review," *Publ. Interest*, fall 1967, 9, 109-18.

- N. Kaldor, "The Economic Aspects of Advertising," *Rev. Econ. Stud.*, No. 1, 1950-51, 18, 1-27.
- D. W. Marquardt, "An Algorithm for Least-Squares Estimation of Nonlinear Parameters," *J. Soc. Ind. Appl. Math.* (SIAM), June 1963, 11, 431-41.
- R. W. Marris, "Galbraith, Solow and The Truth About Corporations," *Publ. Interest*, spring 1968, 11, 37-46.
- M. Nerlove and K. J. Arrow, "Optimal Advertising Under Dynamic Conditions," *Economica*, May 1962, 29, 129-42.
- K. S. Palda, *The Measurement of Cumulative Advertising Effects*, Englewood Cliffs 1964.
- R. M. Solow, (1967a) "The New Industrial State or Son of Affluence," *Publ. Interest*, fall 1967, 9, 100-108.
- , (1967b) "A Rejoinder," *Publ. Interest*, fall 1967, 9, 118-19.
- , "A Comment on Marris," *Publ. Interest*, spring 1968, 11, 47-52.
- G. J. Stigler, "The Economics of Information," *J. Polit. Econ.*, June 1961, 69, 213-25.
- L. G. Telser, "Advertising and Cigarettes," *J. Polit. Econ.*, Oct. 1962, 70, 471-99.
- , "Advertising and Competition," *J. Polit. Econ.*, Dec. 1964, 72, 537-62.
- L. W. Weiss, "Advertising, Profits, and Corporate Taxes," *Rev. Econ. Statist.*, Nov. 1969, 51, 421-30.

# COMMUNICATIONS

## A Note on the Stigler-Kindahl Study of Industrial Prices

By GEORGE A. HAY\*

In a major study recently completed, George Stigler and James Kindahl (S-K) constructed price indexes for various commodity groups based on prices reported by purchasers of the commodity in question. This is in contrast to the Bureau of Labor Statistics (*BLS*) approach in which prices are collected from sellers. The S-K hypothesis is that sellers frequently have an incentive to maintain the stated list price for substantial periods, while at the same time the actual prices at which transactions take place may differ from the list in response to market conditions. The incentives to maintain list price may involve the firm's relations with its competitors, for example, a kinked demand curve notion or the industry's relation with the general public; i.e., if list prices are reduced in periods of insufficient demand, it may be politically difficult to get them back up when conditions change. On the other hand, it is felt that buyers have no such incentive to disguise the purchase price, and prices collected from them should therefore provide a truer picture.

The proper criterion on which to judge a price index good or bad is somewhat ambiguous, since there is no "true" price against which to compare the index. S-K use basically three types of test. The first is a comparison of the trends for the two series. A second test is to plot the two indexes against some measure of general market conditions. S-K use National Bureau reference cycles. If one believes that the true price is correlated

with general business conditions, then a comparison of the correlation of each of the indexes with the National Bureau measures may provide some basis for choosing between the two. A third test is to compare amplitude and frequency of short-run changes in the two indexes.

In addition to being of interest in themselves, however, price indexes are also used as inputs into various econometric models. A relevant criterion for choosing among alternative indexes, therefore, is whether one index significantly improves the explanatory power of a particular regression model.

One such model was developed by the present author in the September 1970 issue of this *Review*. The firm's profit maximizing decisions lead to a set of decision rules for production, price, and finished goods inventory:

$$\begin{aligned} X_t = & A_{11}X_{t-1} + A_{12}P_{t-1} + A_{13}H_{t-1} \\ & + A_{14}U_{t-1} + A_{15}Q_t + A_{16}Q_{t+1} \\ & + A_{17}Q_{t+2} + A_{18}W_{t-1} + A_{19}W_t + k_1 \end{aligned}$$

$$\begin{aligned} P_t = & A_{21}X_{t-1} + A_{22}P_{t-1} + A_{23}H_{t-1} \\ & + A_{24}U_{t-1} + A_{25}Q_t + A_{26}Q_{t+1} \\ & + A_{27}Q_{t+2} + A_{28}W_{t-1} + A_{29}W_t + k_2 \end{aligned}$$

$$\begin{aligned} H_t = & A_{31}X_{t-1} + A_{32}P_{t-1} + A_{33}H_{t-1} \\ & + A_{34}U_{t-1} + A_{35}Q_t + A_{36}Q_{t+1} \\ & + A_{37}Q_{t+2} + A_{38}W_{t-1} + A_{39}W_t + k_3 \end{aligned}$$

where

$X$  = production

$P$  = price

$H$  = finished goods inventory

\* Associate professor of economics at Yale University. Richard Gustafson provided research assistance.

TABLE 1—REGRESSION COEFFICIENTS  
(*t*-values in parentheses)

		$X_{t-1}$	$P_{t-1}$	$H_{t-1}$	$U_{t-1}$	$Q_t$	$Q_{t+1}$	$Q_{t+2}$	$W_{t-1}$	$W_t$
PAPER										
A. S-K Index	1.375	-.016	-.071	-.428	.376	.873	-.056	-.051	7.630	-5.350
$X_t$		(0.52)	(2.17)	(3.90)	(5.95)	(27.06)	(1.79)	(1.79)	(3.80)	(2.56)
	15.454	.165	.877	-.281	.116	-.032	.024	.070	-1.734	-.508
$P_t$		(3.66)	(18.54)	(1.78)	(1.27)	(0.69)	(0.53)	(1.70)	(0.61)	(0.17)
	0.122	.032	-.021	.959	-.050	-.004	.035	.015	4.091	-4.574
$H_t$		(1.61)	(1.00)	(13.67)	(1.23)	(0.20)	(1.72)	(0.85)	(3.26)	(3.44)
B. BLS Index	1.683	-.009	-.078	-.434	.361	.871	-.047	-.063	7.545	-5.086
$X_t$		(0.28)	(3.72)	(3.86)	(5.72)	(26.00)	(1.45)	(2.20)	(3.66)	(2.29)
	16.264	.009	.809	-.082	.277	.058	-.059	.099	-.209	1.799
$P_t$		(0.12)	(15.69)	(2.96)	(1.78)	(0.70)	(0.74)	(1.40)	(0.04)	(0.33)
	-2.166	.032	.024	1.066	-.083	-.012	.044	.002	4.468	-4.931
$H_t$		(1.54)	(1.79)	(14.60)	(2.03)	(0.55)	(2.07)	(0.12)	(3.34)	(3.42)
LUMBER										
A. S-K Index	-1.482	.225	-.034	-.100	.197	.677	.031	.010	.583	.046
$X_t$		(3.82)	(2.22)	(1.01)	(2.86)	(10.04)	(0.45)	(0.18)	(0.35)	(0.03)
	19.910	-.180	.853	-.751	.100	.275	.282	.092	-2.871	-1.118
$P_t$		(1.07)	(19.64)	(2.43)	(0.51)	(1.43)	(1.48)	(0.59)	(0.61)	(0.22)
	-1.000	-.035	.004	.963	.004	-.056	.047	.032	1.919	-1.517
$H_t$		(1.11)	(0.47)	(16.86)	(0.11)	(1.57)	(1.32)	(1.13)	(2.21)	(1.61)
B. BLS Index	-4.618	.159	-.005	-.025	.164	.660	.042	-.029	.835	.318
$X_t$		(2.53)	(0.29)	(0.27)	(2.53)	(9.80)	(0.63)	(0.54)	(0.49)	(0.17)
	10.712	-.139	.883	-.583	.190	.161	.097	.442	-5.428	3.177
$P_t$		(1.05)	(24.36)	(2.94)	(1.39)	(1.13)	(0.68)	(3.86)	(1.52)	(0.81)
	-1.182	-.062	.116	.990	-.008	-.060	.057	.001	2.263	-1.748
$H_t$		(1.88)	(1.75)	(20.04)	(0.22)	(1.68)	(1.61)	(0.05)	(2.53)	(1.79)

 $U$  = backlog of unfilled orders $W$  = average hourly wages<sup>1</sup>

$Q_t$  is intended to represent the expected value of the quantity intercept of a linear demand curve

$$\text{New Orders} \equiv O_t = Q_t - bP_t$$

<sup>1</sup> The theoretical specification in the original model contained a term (denoted as  $V_t$ ) which includes raw materials prices and capital rental as well as labor costs, but for the industries treated here the only series for which there are usable data is the latter.

where  $b$  is the slope which is assumed to remain constant over time. If perfect forecasting is assumed, then  $Q$  is obtained by the inverse relation

$$Q_t = O_t + bP_t$$

For the particular study referred to,  $b$ , which is not observable, was chosen to make the elasticity of demand equal to .5. However, alternative values of  $b$  did not significantly affect the results.

In this model, the BLS index is used in

TABLE 2—COMPARISON OF ALTERNATIVE MEASURES OF GOODNESS OF FIT

Measures	$\bar{R}^2$ <sup>a</sup>	F-Ratio	Std. Error of Estimate $\sigma$	$\sigma +$ Mean of Dependent Variable
I. PAPER				
A. S-K Index				
1. Production	.991	1388.	.019	.0149
2. Price	.985	819.	.276	.0027
3. Inventory	.986	874.	.012	.0226
B. BLS Index				
1. Production	.988	1048.	.020	.0153
2. Price	.830	60.	.500	.0050
3. Inventory	.982	665.	.013	.0238
II. LUMBER				
A. S-K Index				
1. Production	.955	258.	.033	.0445
2. Price	.947	212.	.940	.0091
3. Inventory	.936	177.	.017	.0290
B. BLS Index				
1. Production	.936	180.	.034	.0446
2. Price	.947	219.	.716	.0092
3. Inventory	.910	108.	.018	.0297

<sup>a</sup> 100 degrees of freedom

three ways: 1) as one of the dependent variables; 2) in lagged form as one of the explanatory variables; 3) to deflate the physical variables which are reported by the Census in value terms.

Regressions were performed using both the BLS and the S-K indexes at the two-digit Standard Industrial Classification (SIC) level for the Lumber industry and the Paper industry.<sup>2</sup> These industries are particularly convenient since the commodities covered in the relevant BLS group correspond very closely to the SIC classification used by Census from which the physical data were obtained.<sup>3</sup> The monthly observations are for the period January 1957–August 1966.

<sup>2</sup> While the S-K Price index for paper covers two-thirds of the two-digit Paper Industry weights, their Lumber index covers only 9 percent of the two-digit Lumber Industry weights. Unfortunately the physical data are not available at a lower level of aggregation.

<sup>3</sup> For most industries this is not the case, thereby making it difficult to include price in econometric studies involving production and inventories.

The results of the regression are reported in Table 1. Table 2 attempts to present the various statistics which are commonly used as measures of goodness of fit. It should be noted here that the S-K index is more flexible with standard deviations of 2.277 (mean value 102.95) and 4.018 (mean value 103.12) for Paper and Lumber, respectively, compared with those for the BLS series of 1.209 (mean value 99.96) and 3.112 (mean value 99.42).

No attempt will be made to interpret the individual coefficients (see Hay). The main result worth noting here is that there are no significant sign differences between the two sets and almost all the coefficients display the same order of magnitude with both price variables.

The interesting results are in Table 2. There it is clear that the S-K index improves the fit of the model. Except for the price equation in Lumber in which the BLS index does slightly better, the S-K index comes out ahead under every measure for every equation:  $\bar{R}^2$  values and F-ratios are higher, standard errors and standard errors divided by the mean of the corresponding dependent variable are lower. In general, the improvements are not large, however, and the evidence can certainly not be considered conclusive.

It would be desirable to repeat the experiment for a greater number of industries (however, see footnote 2) and to make comparisons over other time periods. However, the results do suggest that the S-K index may be an improvement not only as an index per se, but also as an input into econometric models which require a price variable either as a deflator or as one of the variables.

## REFERENCES

- G. A. Hay, "Production, Price and Inventory Theory," *Amer. Econ. Rev.*, Sept. 1970, 60, 531–45.  
G. J. Stigler and J. K. Kindahl, *The Behavior of Industrial Prices*, New York 1970.

# The Statistical Theory of Racism and Sexism

By EDMUND S. PHELPS\*

My recent book, *Inflation Policy and Unemployment Theory*, introduces what is called the statistical theory of racial (and sexual) discrimination in the labor market.<sup>1</sup> The theory fell naturally out of the non-Walrasian treatment there of the labor "market" as operating imperfectly because of the scarcity of information about the existence and characteristics of workers and jobs.

A paradigm for the theory is the traveller in a strange town faced with choosing between dinner at the hotel and dinner somewhere in the town. If he makes it a rule to dine outside the hotel without any prior investigation, he is said to be discriminating against the hotel. Though there will be instances where the hotel cuisine would have been preferable, the rule represents rational behavior—it maximizes expected utility—if the cost of acquiring evaluations of restaurants is sufficiently high and if the hotel restaurant is believed to be inferior at least half the time.

In the same way, the employer who seeks to maximize expected profit will discriminate against blacks or women if he believes them to be less qualified, reliable, long-term, etc. on the average than whites and men, respectively, and if the cost of gaining information about the individual applicants is excessive. Skin color or sex is taken as a proxy for relevant data not sampled. The a priori belief in the probable preferability of a white or a male over a black or female candidate who is not known to differ in other respects might stem from the employer's previous statistical experience with the two groups (members from the less favored groups might have been, and continue to be, hired at less favorable terms); or it might stem from prevailing

sociological beliefs that blacks and women grow up disadvantaged due to racial hostility or at least prejudices toward them in the society (in which latter case the discrimination is self-perpetuating).

The theory is applicable to the class of "liberal" employers and workers who have no distaste for hiring and working alongside black or female workers. By contrast, the theory of discrimination originated by Gary Becker is based on the factor of racial taste. The pioneering work of Gunnar Myrdal et al. also appears to center on racial (and, in an appendix, sexual) antagonism.

Some indications of interest in the new theory, and the independent discovery of the same statistical theory by Kenneth Arrow, convince me that it is time for a formalization of the theory in terms of an exact statistical model. Though what follows is very simple, it may be useful to those who like exact models and it may stimulate others to develop the theory further.

An employer samples from a population of job applicants. The employer is able to measure the performance of each applicant in some kind of test,  $y_i$ , which, after suitable scaling, may be said to measure the applicant's promise or degree of qualification,  $q_i$ , plus an error term,  $\mu_i$ .

$$(1) \quad y_i = q_i + \mu_i$$

where  $\mu$  is normally distributed with mean zero.

It is conceivable (and it sometimes occurs in practice) that the employer will have no other information about each applicant, including skin color.<sup>2</sup> In that special case, the employer may use  $q_i$  as a least-squares predictor of the applicant's  $y_i$  according to the regression-type relation:

\* Professor of economics, Columbia University. The paper was written under a grant from the Fels Institute, University of Pennsylvania.

<sup>1</sup> I am indebted to Edward Prescott and Karl Shell for proposing the extension of the paper to Case 2.

<sup>2</sup> The Fair Employment Practices Law forbids employers from asking for information on race in written applications. The Boston Symphony Orchestra auditions candidates from behind an opaque screen.

$$(2) \quad q'_i = a_1 y'_i + u'_i$$

$$0 < a_1 = \frac{\text{var } q'_i}{\text{var } q'_i + \text{var } \mu_i} < 1, E u_i = 0$$

where  $q'_i$  and  $y'_i$  are deviations from their respective population means.<sup>3</sup>

Suppose instead that skin color is observed along with the test datum, and suppose that the employer postulates a model of job qualification

$$(3) \quad q_i = \alpha + x_i + \eta_i$$

in which

$$(3a) \quad x_i = (-\beta + \epsilon_i)c_i, \quad \beta > 0,$$

where  $c_i = 1$  if the applicant is black and zero otherwise. Here  $x_i$  is the contribution of social factors, and these are believed to be race-related according to (3a). The random variables  $\epsilon_i$  and  $\eta_i$  are normally and independently distributed with mean zero. Letting  $\lambda_i = \eta_i + c_i \epsilon_i$  and  $z_i = -\beta c_i$ , we may write

$$(4) \quad q_i = \alpha + z_i + \lambda_i$$

$$y_i = q_i + \mu_i = \alpha + z_i + \lambda_i + \mu_i$$

Then the test datum can be used in relation to the race (sex) factor to predict the degree of qualification net of the race factor, the latter being separately calculable:

$$(5) \quad q'_i - z'_i = a_1 \cdot (y'_i - z'_i) + u_i$$

$$0 < a_1 = \frac{\text{var } \lambda}{\text{var } \lambda + \text{var } \mu} < 1$$

or, equivalently

$$(5') \quad q'_i = \frac{\text{var } \lambda}{\text{var } \lambda + \text{var } \mu_i} \cdot y'_i + \frac{\text{var } \mu_i}{\text{var } \lambda + \text{var } \mu_i} \cdot z'_i + u_i$$

<sup>3</sup> In (2),  $a_1$  is the probability limit, as  $N \rightarrow \infty$ , of the regression coefficient

$$\hat{a}_1 = \frac{\frac{1}{N} \sum_{i=1}^N y'_i q'_i}{\frac{1}{N} \sum_{i=1}^N (y'_i)^2}$$

$$= \frac{\frac{1}{N} \sum_{i=1}^N (q'_i + \mu_i) q'_i}{\frac{1}{N} \sum_{i=1}^N (q'_i + \mu_i)^2}$$

The weights applied to the test information and the skin color information are inversely related to the variances of the respective disturbance terms corresponding to them.<sup>4</sup>

CASE 1. If growing up black is believed by the employer to be socially disadvantageous, so that  $z'_i < 0$  for black applicants, then one might expect to find a lower prediction of  $q_i$  for blacks than whites having equal test scores. This is generally true, however, only in the special case where  $\epsilon_i \equiv 0$  for all  $i$ , i.e., for all blacks as well as whites. This means that there is no differential variability in promise as between blacks and whites. Then  $\text{var } \lambda_i = \text{var } \eta_i$  and hence the coefficients in (5') are independent of  $c_i$ . Therefore the prediction curve relating  $q_i$  to  $y_i$  for blacks lies parallel and below that for whites, as illustrated in Figure 1.

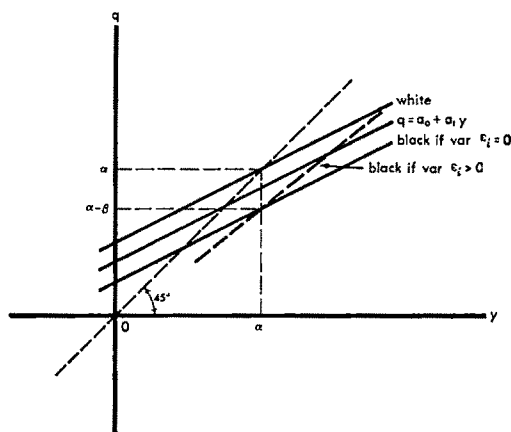


FIGURE 1. PREDICTION OF QUALIFICATION BY RACE AND TEST SCORE

CASE 2. In general the variance of  $\lambda$  depends upon skin color. The formulation in (3) ascribes to blacks the larger postulated variance, as reflected in (6):

$$(6) \quad \text{var } \lambda_i = \text{var } \eta_i + c_i^2 \text{var } \epsilon_i$$

<sup>4</sup> My attention has been called by the referee to the derivation of a generalization of equation (5'), from which can be deduced all my cases, in the extended footnote on page 325 in Thomas Wonnacott and Ronald Wonnacott.

It follows that the coefficient of the test score in the least-squares prediction of qualification is *greater* for blacks than for whites. (In the limit, as  $\text{var } \epsilon_i \rightarrow \infty$ , the coefficient of  $y_i$ —the slope of the prediction curve for blacks—approaches one.) For any positive  $\text{var } \epsilon_i$  it is a consequence of the race-related difference in coefficients that at some high test score and higher ones the black applicant is predicted by the employer to excel over any white applicant with the same or lower score. The employer credits an equally good test score by the white applicant as a *less credible* indication in view of the prior notions of the comparatively narrow range of white promise. Note that one can reverse these implications by replacing the dummy variable in (3) with  $(1 - c_i)$  instead.

A FURTHER CASE. It is straightforward to make the disturbance term in (1) conditional on race in the way that  $\lambda$  was made conditional on skin color:

$$(7) \quad \mu_i = \xi_i + c_i p_i$$

Then whites' test scores are regarded by the employer as *more reliable* than the scores of blacks—that is, they measure promise with less error. In that case the greater reliability of whites' test scores might overcome any tendency for them to have less credi-

bility, so that the white prediction curve would be the steeper curve. Then there is a range of low test scores in which whites are predicted to be less qualified than equally high scoring blacks.

A final word. A sensitive person, I have been warned, might read this paper as expressing an impression on the part of the author that most or all discrimination is the result of beliefs that blacks and women deliver on the average an inferior performance. Actually, I do not know (nor claim to know) whether in fact most discrimination is of the statistical kind studied here. But what if it were? Discrimination is no less damaging to its victims for being statistical. And it is no less important for social policy to counter.

#### REFERENCES

- K. J. Arrow, "Some Models of Racial Discrimination in the Labor Market," RAND Corporation research memorandum RM-6253-RC, multilith, Santa Monica, Feb. 1971.
- G. S. Becker, *The Economics of Discrimination*, Chicago 1959.
- G. Myrdal, *An American Dilemma*, New York 1944.
- E. S. Phelps, *Inflation Policy and Unemployment Theory*, New York 1972.
- T. H. Wonnacott and R. J. Wonnacott, *Introductory Statistics*, New York 1969.

# Learning and Productivity Change in Metal Products

By LEONARD DUDLEY\*

This paper tries to relate the process of technical change to the characteristics of the production processes in a particular sector of a developing economy. The example chosen is the metal-products sector of Colombia, where over a seven-year period from 1959 to 1966 real output per worker increased by 6 percent annually. It is argued that this productivity growth may be explained best by a hypothesis of learning from production experience. Such a hypothesis, however, should take account of interindustry differences both in the process of learning (cumulative output versus elapsed time) and the learning agent (firm<sup>1</sup> versus worker).

## I. Alternative Explanations of Productivity Change

There are available in the economic literature a number of explanations of productivity change which might apply to the Colombian metal products sector.

One possibility is a neoclassical production function,

$$(1) \quad \pi = f(k, t, s)$$

in which output per workers,  $\pi$ , is a function of capital per worker,  $k$ , disembodied technical change occurring at a uniform annual rate,  $t$ , and the scale of the firm or the industry,  $s$ .

Richard Nelson has argued, however, that instead of a single production function there may at any point in time be different types of manufacturing firms, between which are gaps in technology. To state the hypothesis in simplified form for one group, the modern firms, the production function may well be of the type expressed in equation (1). For the

other group, the craft firms, the production function is of the simplest neoclassical type, with no technical change or scale economies.<sup>2</sup> Letting  $m$  and  $c$  represent modern and craft, respectively:

$$(2a) \quad \pi_m = f_m(k_m, t, s_m)$$

$$(2b) \quad \pi_c = f_c(k_c)$$

where at the same capital-labor ratio  $\pi_c < \pi_m$ . Under these dualism assumptions, the observed productivity increases in Colombian metal products would be interpreted as the consequence of expansion in the modern sector at the expense of the craft sector.<sup>3</sup>

A third hypothesis is that at a constant capital-labor ratio and constant scale, productivity may increase substantially as production experience,  $G$ , increases:

$$(3) \quad \pi = f(k, s, G)$$

This learning hypothesis has the advantages both of being consistent with observed experience and of being capable of explaining substantial increases in productivity over time.<sup>4</sup> Still, there remains the problem of

<sup>2</sup> Note that this technological dualism, characterized by *different production functions* for large and small firms is different from the type of dualism that can occur under the neoclassical version if small and large firms have the same production function but face *different factor prices*.

<sup>3</sup> E. Mansfield and Richard Nelson, and Karsten Laursen and Lester Taylor have used this method to explain productivity differences between small and large firms for Colombian two-digit industries.

<sup>4</sup> Various measures of  $G$  have been suggested. A relationship between productivity and cumulative investment has been proposed by Kenneth Arrow and tested by Eytan Sheshinski. And a simple relationship between productivity and time has been proposed by William Fellner and Robert Zevin with empirical support from Paul David. However, the most commonly used measure of this production experience is cumulative output. Studies of various metal products industries by H. Asher, Werner Hirsch and Leonard Rapping have indicated that a doubling of cumulative output may yield a productivity increase of from 10 to 30 percent, most of the gain coming from a reduction in unit labor requirements.

\* Université de Montréal. This paper is based on a doctoral dissertation presented to Yale University. I wish to thank Richard Nelson, R. Albert Berry, Howard Pack, and David Grether for their helpful comments and suggestions.

<sup>1</sup> Learning by firms is a concept explained below in Section II.

selecting the most appropriate measure of *G*, to which the next section turns.

## II. A Specification of the Learning Hypothesis for Metal Products Characteristics of Metal Products Production

The production activities carried out in the Colombian metal products sector may be divided roughly into four groups: casting, forging, and stamping (*CFS*); metal working; assembly; and repair. To distinguish among these branches, it is useful to specify several characteristics of the design and production processes.

One of these dimensions is the complexity of the production task of each worker, as defined perhaps by the number of substeps in the task and the skills required in each substep. By this criterion the metal working activity probably has the highest ranking. The other branches are nevertheless characterized by an intermediate degree of complexity, as shown in Table 1.

A second characteristic is the degree of sequential interdependence among the tasks necessary to finish the product. This interdependence, which might be measured by the total number of tasks in the production process, is greatest in assembly operations.

Third, one should consider the importance of design and development operations in the activity, as measured by the percentage of the labour force consisting of nonproduction employees. This characteristic is most apparent in metal working, since it is at this stage that design specifications must be translated into production operations.

### *Who Learns: Workers or Firms?*

Given these production characteristics, what can be said about the nature of the learning process in each activity. For example, who learns? The simplest hypothesis is that learning is embodied in individual workers. If, as seems reasonable, the importance of such worker-embodied learning is a function of the complexity of the task involved, it should be most evident in metal working.

In some activities, however, learning by workers may be less important than learning that becomes embodied in the production system of the firm itself—that is, learning by

TABLE 1—PRINCIPAL CHARACTERISTICS OF METAL PRODUCTS PRODUCTION BY ACTIVITY

	Casting, Forging, Stamping	Metal Working	Assembly	Repair
<i>Characteristic</i>				
Production-task complexity	int. <sup>a</sup> -high	high	int.	int.
Production-task interdependence	int.	low	high	low
Importance of non-production tasks	int.	high	low	low
<i>Who learns?</i>				
a. Workers	int.-high	high	int.	int.
b. Firms:				
Production				
Coordination	int.	low	high	low
Production-nonproduction coordination	int.	high	low	low
<i>How?</i>				
a. Workers:				
Output-dept	int.	low	int.	int.
Time-dept	int.	high	low	low
b. Firms:				
Output-dept	int.	low	high	low
Time-dept	int.	high	low	low

<sup>a</sup> intermediate.

the firm. With experience, a firm may develop a system for training workers and managers, designing products, and coordinating operations of suppliers with its own. The result is a package relative to which the experience of individual workers and managers is an easily acquired and unimportant factor.

### *How Does Learning Occur? Output versus Time*

The other question to be answered is how this learning occurs. There are two principal alternatives, one suggesting that learning depends on cumulative output and the other that it depends on elapsed time.<sup>5</sup> However,

<sup>5</sup> See fn. 4. As Richard Nelson remarked, the Arrow hypothesis that learning is a function of *cumulative investment*, taken in the context of the one-sector model for which it was proposed, is really equivalent to learning as a function of the *cumulative output* of the capital goods industry.

the studies of Hirsch and Fellner suggest that these hypotheses are not necessarily mutually exclusive.<sup>6</sup> On the one hand, it is plausible that learning through repetition of the same physical task, which might be called *simple learning*, is closely related to the level of cumulative output. On the other hand, there may be a second process of learning to learn which leads to an increase in this simple learning rate over time. For various reasons, this process may depend less on past production than on time elapsed.<sup>7</sup>

Consider now the learning that is embodied in individual workers. Where the tasks are relatively simple (assembly, repair), this learning is likely to be output-dependent. Where more complex operations are involved (*CFS*, metal working) labor embodied learning is probably time-dependent.

When it comes to learning by firms, the production-coordination aspect would appear to be a simpler type of learning, akin to the acquisition of production skills. In assembly, then, firm learning is likely to depend on cumulative output. In contrast, the

<sup>6</sup> Hirsch has offered evidence which indicates that there may be two types of learning occurring. On a certain metal products job, production of lathes of a certain type, he observed a "progress ratio" of 18; that is, there was an 18 percent decline in unit labor requirements associated with a doubling of cumulative output. However, this rate of productivity change itself rose with production experience; the next job, to produce lathes of a different type, had an associated progress ratio of 25. Thus, in addition to a simple learning process to explain productivity changes on a single job, there may be a separate learning to learn process to explain an increase in the progress ratio between jobs. It may be possible to distinguish further between these two types of learning. Fellner's study of Olympic sports found that in simpler sports, where there had been little change in equipment or rules, performance was closely related to cumulative output. In other cases, where strategies were more complex or where rules or equipment had been modified, performance was explained better as a function of time.

<sup>7</sup> Owing to model changes over time, a given volume of cumulative output is likely to involve a larger number of different jobs and therefore a larger variety of experience, the longer the time period over which that output is spread. In addition, it might be suggested that in any complex cognitive process, such as learning how to acquire production or athletic skills, time may be the key explanatory variable.

more complex coordination of development and production tasks, which is important in metal working, should be primarily time-dependent.

Thus several interesting hypotheses are suggested:

1. In the metal products sector a production function based on learning should dominate the neoclassical and technological dualism versions.

2. It may be possible to distinguish between two types of learning; one a simple learning process dependent on cumulative output, the other a process of learning to learn, dependent on elapsed time.

3. It may also be possible to distinguish between two types of learners—firms and their workers.

### III. The Importance of Learning

#### *The Model*

The model used to test the first of these hypotheses was a single-equation Cobb-Douglas production function estimated by ordinary least squares. In order to distinguish between the neoclassical approach and learning by doing, it was necessary to formulate different versions of the model.

Corresponding to the neoclassical equation (1) is the version:

$$(4) \quad \ln (V/L)_{jt} = a_0 + \alpha_0 \ln (K/L)_{jt} + \gamma_0 \ln L_{jt} + \lambda_0 t + w_{jt}$$

where  $V$  is value-added,  $K$  is the energy capacity in horsepower,  $L$  is the number of employees,  $j$  indicates the size class,  $t$  the year, and  $w$  is a disturbance.

A second version of the model permits learning, following equation (3).

$$(5) \quad \ln (V/L)_{jt} = a_i + \alpha_i \ln (K/L)_{jt} + \gamma_i \ln L_{jt} + \lambda_i t + \eta_i G_{irt} + w_{ijt} \\ i = 1, 2, 3, 4$$

where

$$r = 1, 1 \leq j \leq 6 \\ r = 2, 7 \leq j \leq 10$$

Here  $G_i$  is a measure of learning, where the

subscript  $i$  indicates the learning regime, as described below. The subscript  $r$  indicates the size group.

These functions were fitted separately to each of twenty-five Colombian three-digit metal products industries, using pooled cross-section time-series data. There were ten size classes of firms (by the number of employees per firm), each size class having a maximum of eight observations, one for each of the years 1959–66. It should be noted that it proved impossible to obtain measures of  $G_i$  for each of the ten size classes. Consequently, estimates were made for small firms (1–49 employees,  $r=1$ ) on the one hand, and for large firms (50 employees and over,  $r=2$ ) on the other. In addition, all monetary values were deflated by the same price index for home and imported goods.

#### *Indices of Learning*

From the discussion of Section II, it is evident that there is a four-way breakdown of possible learning processes, as shown in Table 2. In each case, a possible measure of learning suggests itself: the logarithm of cumulative output per worker and per firm for the output-dependent type; average years of experience per worker and average age of firm for the time-dependent type. Although data were not available at the firm level, it was possible to construct indices which approximate the desired measures.<sup>8</sup>

#### *The Statistical Significance of Learning*

Of the five models described in equations (4) and (5), only the learning model with average years of experience per worker  $G_3$  failed to produce significant results. Table 3

<sup>8</sup> To obtain these indices, a base-year (1959) value was first assigned fairly arbitrarily to each index. The data for each succeeding year were then used to modify the index. For cumulative value-added, this modification consisted simply of adding the current year's value added to the cumulative total. For the age-of-firm and years of experience variables, first differences between the total number of firms or workers in successive years were assumed to constitute new additions to the stock of firms or workers. To handle movements between the two size classes, it was assumed that all new entrants to the industry start as small firms (under 50 employees). Thus changes in the number of large firms were assumed caused by movements between size classes.

TABLE 2—FOUR TYPES OF LEARNING IN METAL PRODUCTS AND A POSSIBLE MEASURE OF EACH TYPE

Type of Learning	Measure
1. Learning by workers dependent upon cumulative output	Log of cumulative value-added per employee ( $G_1$ )
2. Learning by firm dependent upon cumulative output	Log of cumulative value-added per establishment ( $G_2$ )
3. Learning by workers dependent upon elapsed time	Average number of years of experience per employee ( $G_3$ )
4. Learning by firm dependent upon elapsed time.	Average age of establishment ( $G_4$ )

compares the neoclassical version with the best fitting of the three remaining learning versions.

On the whole, the results indicate that learning was statistically significant in Colombian metal products. By the  $t$ -test, the cumulative output per establishment variable was significant at the .05 level in fifteen out of twenty-five industries, while the two remaining learning versions were significant in twelve and nine industries, respectively. The noticeably better fit of the learning versions compared with the neoclassical version indicates that the learning variables were not simply substituting for the effect of the time variable in the latter version.

The addition of the learning variables had an interesting effect on  $\lambda$ , the time trend. In many cases where  $\lambda$  had been nonsignificant or even significant with a positive sign in the neoclassical version, the coefficient appeared significant with a negative sign in the learning versions. It would seem, therefore, that the learning effect may be a compound one, made up of the influences of both the learning variable and the time variable.

#### *The Economic Importance of Learning*

When one turns to consider the economic importance of learning, interpretation of the statistical results becomes somewhat more difficult. What is necessary is a way to compare the effects of the various causes of productivity change. One may do this by combining the elasticities of Table 3 with the annual rates of growth in the corresponding

TABLE 3—COEFFICIENTS OF COBB-DOUGLAS PRODUCTION FUNCTIONS FOR COLOMBIAN METAL PRODUCTS, WITH AND WITHOUT LEARNING, FOR TWENTY-FIVE THREE-DIGIT INDUSTRIES, 1959-66

Industry	No learning			$F^2$	With learning (using measure $G_2$ ) <sup>a</sup>				$R^2$
	$\alpha_0$	$\gamma_0$	$\lambda_0$		$\alpha_2$	$\gamma_2$	$\lambda_2$	$\eta_2$	
Tinware products	.33*	.21*	.05*	.49	.30*	.12*	.04*	.13*	.55
Tools	.44*	.29*	-.02	.80	.42*	.27*	-.04	.07	.80
Cutlery	1.19*	.13	.06	.59	1.21*	.16	.07	.03	.59
Nonelectrical appliances	.22*	.12*	-.03	.38	.23	.09	-.03	.05	.29
Aluminum products	.28*	.10*	.00	.34	.12*	.04	-.03	.14*	.48
Wire products	.32*	.14*	.00	.43	.29*	.12*	.00	.10	.44
Foundries	.62*	.17*	-.01	.62	.58*	.12	-.02	.10*	.65
Machine shops	.18*	.18*	.01	.33	.09	.05	.01	.15*	.49
Other metal products	.48*	.10	.03	.35	.45*	.01	.02	.08	.37
Turbines	.24	.18*	-.05	.8	.25	.03	-.07	.25*	.26
Agric. machinery	-.23*	.30*	.04*	.36	-.10	.23*	.01	.19*	.55
Industrial machinery	-.62*	-.11	.07	.0	-.62*	-.16	.06	.07	.11
Machinery parts	.07	.03	.06*	.7	.19*	.03	.02	.16*	.34
Electrical machinery	.10	.29*	.02	.16	.14*	.18*	-.03*	.09*	.68
Radio and television	.07	.39*	-.05*	.31	.07	.37*	-.05*	.01*	.51
Electric appliances	-.36*	.22*	.08*	.37	-.32*	.06	.07*	.19*	.45
Electric wire	.38*	.17*	-.03	.37	.35*	.00	-.04	.18*	.63
Light bulbs	.01	.12*	.00	.06	.00	.10	-.01	.10*	.07
Electric installations	.18	.46*	-.04	.48	.33*	.34*	.05*	.19*	.53
Naval repair	.04	.25*	.00	.39	.05	.24*	-.00	.01	.39
Railway repair	.01	.12*	.09*	.54	.02	.11*	-.09	1.12	.70
Automobiles	.53*	.10	-.02	.37	.54*	.05	-.03	.08	.38
Bicycles	.17*	.27*	-.01	.47	.22*	.08	-.07*	.31*	.65
Automobile repair	-.28*	-.19*	-.01	.12	-.22*	-.16*	-.02	.05*	.45
Aircraft repair	.09*	.07*	.01	.21	.08*	.01	.01	.08*	.29
Averages:									
value of coefficient	.18	.17	.01	.40	.19	.10	-.01	.15	.47
t-statistic	(2.8)	(3.5)	(1.2)		(2.9)	(1.9)	(1.0)	(2.4)	

<sup>a</sup> Log of cumulative value-added per establishment.

\* Coefficient significant at .05 level.

explanatory variables,  $k$ ,  $L$  and  $G_i$  ( $i=1, 2, 4$ ). Table 4 summarizes the resulting breakdown of total productivity change.<sup>9</sup>

Whichever measure of learning is chosen, the effect of learning would appear to be considerably more important than the effects of increased capital per worker or larger scale. By itself, learning would seem capable of explaining annual productivity increases of from 2 to 3 percent in the sector as a whole, with considerably higher rates in individual industries.

<sup>9</sup> Note that estimates were made only for those coefficients of Table 3 that were statistically significant from zero at the .05 level of significance.

### *Learning and/or Dualism*

Despite the apparent superiority of the learning version over the neoclassical version, it is conceivable that the productivity changes in Colombian metal products may be explained equally well by a Nelson-type hypothesis that small and large firms have different neoclassical production functions.

If the neoclassical version is used to test for dualism, the Chow test leads to rejection of the hypothesis of identical production functions at the .01 level in fifteen of twenty-four industries. The metal products group would therefore appear to be characterized by structural dualism. In the learning ver-

TABLE 4—ESTIMATED BREAKDOWN OF PRODUCTIVITY CHANGES IN COLOMBIAN METAL PRODUCTS, 1959–1966 (AVERAGES FOR TWENTY-FIVE THREE-DIGIT INDUSTRIES)

Version estimated	Estimated percent change in productivity due to:				Total
	<i>K/L</i>	<i>L</i>	<i>t</i>	<i>G<sub>i</sub></i>	
Without learning (equation (4))	9	11	10	—	30
With learning (equation (5)) using:					
<i>G<sub>1</sub></i>	9	9	—5	31	44
<i>G<sub>2</sub></i>	9	6	2	12	29
<i>G<sub>4</sub></i>	10	9	—18	34	35
Actual productivity change, 1959–66					51

sion, however, the hypothesis of identical functions is rejected in only nine of the twenty-four cases using cumulative value-added per establishment (ten or twelve cases using cumulative value-added per employee or average age of establishment, respectively).

Thus productivity differences between small and large firms would appear to arise as much from differences in experience levels of firms and their workers as from differences in the technology being used.

#### IV. Types of Learning

Nothing has yet been said about the second and third hypotheses, which refer to the characteristics of the learning process in metal products. An intriguing question is whether one may compare the importance of the different types of learning in each industry, simply by examining the learning effects shown in Table 4. The problem with such a procedure is that it almost certainly involves some double-counting of learning. Because of multicollinearity between *G<sub>1</sub>*, *G<sub>2</sub>* and *G<sub>4</sub>*, the figure corresponding to *G<sub>1</sub>* will include part of the effects of *G<sub>2</sub>* and *G<sub>4</sub>*. Nevertheless, while this separate estimation procedure will overestimate the *magnitude* of the effect of each of the three types of learning, it will still leave the *ranking* of any two effects unchanged.<sup>10</sup>

<sup>10</sup> Assume that

$$G_1 = X + u$$

$$G_2 = X + Y + u$$

Two additional steps are necessary to be able to distinguish between types of learning by activity. In the first step, the effects of time and learning are added together for the sixteen industries in which one or more of the learning indices produced significant coefficients. These industries may be grouped into four categories, corresponding to the relevant two-digit industries (except that the bicycle industry is grouped with the similar industries of the nonelectrical machinery group).

The second step occurs in Table 5. It will be remembered that two of the learning indices were based on cumulative value-added. In columns 1 and 2 of this table, the average effect of these two indices is compared with that of the average age of establishment. In this way, the effect of cumulative value-added may be compared with the effect of elapsed time in each of the four groups. In the heterogeneous metal products group, cumulative value-added and elapsed time were both important, although elapsed time had a slight edge. In nonelectrical machinery and bicycles, where metal working activities are combined with assembly, cumulative value-added was still important,

where *X* and *Y* are both positive and independently distributed and *u* is a disturbance. Note that because both learning indices are a function of *X*, there will be high multicollinearity between them. Suppose that the true relationship between productivity and learning is an exact one:  $p = a + \eta(X + Y)$ . Thus *G<sub>1</sub>* embodies only a part of the total learning effect. The problem is to see whether the resulting estimate of  $\eta$  is smaller than that when *G<sub>2</sub>* which embodies the whole learning effect is used. If the relationship that is estimated is  $p = \hat{a}_1 + \hat{\eta}_1 G_1 + e_1$ , it may be shown that the limiting value to which  $\hat{\eta}_1$  tends in probability as the number of observations increases is

$$p \lim \hat{\eta}_1 = \frac{\eta}{1 + \sigma_u^2 / \sigma_x^2}$$

Similarly if the equation is estimated in the form  $p = \hat{a}_2 + \hat{\eta}_2 G_2 + e_2$  the corresponding estimate is

$$p \lim \hat{\eta}_2 = \frac{\eta}{1 + \sigma_u^2 / \sigma_{X+Y}^2}$$

Since *X* and *Y* are independent,

$$\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2 > \sigma_X^2$$

Therefore  $p \lim \hat{\eta}_1 < p \lim \hat{\eta}_2 < \eta$ .

TABLE 5—COMPARISON OF THE EFFECT ON LABOR PRODUCTIVITY OF FOUR TYPES OF LEARNING, 1959–1966 (FOUR COLOMBIAN METAL PRODUCTS GROUPS)

Group	Estimated percent change in productivity due to:			
	Cum. Value-Added (1)	Elapsed Time (2)	Learning by Workers (3)	Learning by Firms (4)
Metal products	26	38	37	26
Nonelectrical machinery, bicycles	22	48	18	37
Electrical machinery	34	5	35	19
Repair	14	0	20	4

as might be expected. But here, elapsed time appears to be the dominant influence on productivity. In contrast, the electrical machinery group, which is predominantly an assembly activity, received only a small productivity increase from elapsed time. Cumulative value-added, however, had a major effect on productivity in this group. In repair, cumulative value-added also had the more important effect, though this effect was smaller than in the other groups.

These results, tenuous though they are, do seem consistent with the second hypothesis. For activities with production and managerial tasks of considerable complexity—as in metal working—productivity seems to be primarily a function of elapsed time. For simpler production and managerial operations—as in assembly and repair—productivity seems to a great extent to be a function of cumulative value-added.

Less illuminating was a similar comparison of learning by workers (using cumulative value-added per employee) with learning by firms (using the average of the effects of cumulative value-added per establishment and average age of establishment) in columns 3 and 4 of Table 5. Firm learning proved to be important in all branches except repair, as the third hypothesis suggests. Learning by workers was important in all groups—also in accord with the hypothesis. However contrary to expectations, it was least important in nonelectrical machinery and bicycles,

in which the metal working activity accounts for a large part of total value-added.

The relative unimportance of labor-embodied learning in nonelectrical machinery may perhaps be explained by a failure to obtain significant results for the labor-experience version of the learning model. Section II argued that worker learning in metal working would be primarily a function of time, due to the complexity of the production tasks. Since the variable measuring average years of experience per employee did not prove significant, the calculations of Table 5 make no allowance for labor-embodied learning.<sup>11</sup>

Thus the empirical results appear consistent with hypotheses 2 and 3. Both the nature of the learning process and the learning agent differ between production activities.

## V. Conclusion

The implications of the learning version of industrial development differ substantially from those of the versions emphasizing capital, scale, or technology. These more conventional versions all imply the need for a developing economy to remain open, welcoming foreign investment and foreign aid, and thriving in the rigors of international competition. But the answer is not indiscriminate protection of metal products industries. Such a policy is likely to founder because of short-

<sup>11</sup> But the results with the labor-experience variable may be interesting precisely because of their nonsignificance. As defined in Section III, this variable would measure average years of experience of employees only if mobility of experienced workers were low. It was assumed in the computation of the index that all additions to the labor force in a size group consist of inexperienced workers. If, in fact, there was considerable recruitment of experienced workers, the index will underestimate the average experience level of workers. Moreover, if the extent of such recruitment is a function of firm size, the degree of underestimation will not be uniform across size groups. Sligh, pp. 57–58, has observed that during this period of oversupply on the labor market, large high-wage firms did in fact try to use their hiring policies to improve the quality of their labor force. If so, the index used would considerably underestimate the experience level of workers in large firms. The nonsignificance of this variable is at least consistent with the existence of significant mobility of experienced workers in metal products industries.

ages of precisely those skills it might be thought to create. Although to a certain extent these obstacles may be bypassed through the importation of skill-intensive intermediate inputs, eventually a program of import substitution reaches a point at which further import substitution is prohibitively expensive.

What this study implies is that the import substitution barrier in metal products may be breached—at a price. The learning process in metal products requires both time and production experience to enable firms and their workers to acquire the necessary production and managerial skills. If firms fail to internalize a significant part of the benefits of learning, they should be subsidized during the learning period. The results of the study offer some guidelines for an optimal strategy of subsidization in metal products.

With regard to the question of whom to subsidize, metal working and, to a lesser extent, casting, forging, and stamping are the principal candidates, since external economies from learning are probably highest in these activities. Contrary to current practice, heavy protection is less desirable in assembly, where externalities from learning are probably lower.

As for the question of how to subsidize, this study has suggested that time is equally as important for learning as cumulative output—especially in industries such as metal working, where the production process is complex. If so, these subsidies should not be designed solely to encourage a high volume of output in as short a time as possible, but should try to spread this output over time. One means of subsidizing both experience in terms of time, and experience in terms of cumulative output might be per-unit subsidies up to a certain volume of output, with no subsidies beyond this level.

The study has also suggested that learning by firms (in the sense of the development of a system for controlling production and non-

production activities) may be as important as learning by workers. If so, it may be possible to speed up the learning process by engaging in joint ventures with foreign capital or by borrowing foreign technical assistance.

#### REFERENCES

- K. J. Arrow, "The Economic Implications of Learning by Doing," *Rev. Econ. Stud.*, June 1962, 39, 155-73.
- H. Asher, *Cost-Quantity Relationships in the Airframe Industry*, Santa Monica 1956.
- P. A. David, "Learning by Doing and Tariff Protection: A Reconsideration of the Case of the Antebellum U.S. Cotton Textile Industry," *J. Econ. Hist.*, Sept. 1970, 30, 521-601.
- W. Fellner, "Specific Interpretations of Learning by Doing," *J. Econ. Theor.*, Aug. 1969, 1, 119-40.
- W. Z. Hirsch, "Firm Progress Ratios," *Econometrica*, Apr. 1956, 24, 136-43.
- K. Laursen and L. D. Taylor, "Unemployment, Productivity and Growth in Colombia," mimeo., Bogotá 1968.
- E. Mansfield and R. R. Nelson, *Production Functions for a Dual Industrial Structure—Colombian Manufacturing*, Santa Monica 1968.
- R. R. Nelson, "A Diffusion Model of International Productivity Differences in Manufacturing Industry," *Amer. Econ. Rev.*, Dec. 1968, 58, 1219-48.
- L. Rapping, "Learning in World War II Production Functions," *Rev. Econ. Statist.*, Feb. 1965, 47, 81-86.
- E. Sheshinski, "Tests of the Learning by Doing Hypothesis," *Rev. Econ. Statist.*, Nov. 1967, 49, 568-78.
- R. L. Slighton, *Relative Wages, Skill Shortages, and Changes in Income Distribution in Colombia*, Santa Monica Oct. 1968.
- R. B. Zevin, Unpublished paper on long-run learning in a U.S. Antebellum Cotton Firm, 1969, referenced in article by P. A. David.

# The Number of Firms and Competition

By EUGENE F. FAMA AND ARTHUR B. LAFFER\*

It is common economic doctrine that, strictly speaking, with less than an infinite number of firms in an industry, the demand curve facing any firm is negatively sloped. Moreover, the degree to which a firm faces a less than perfectly elastic demand curve is presumed to depend in part on the number of firms, with perfect competition arising in the limit as the number of firms approaches infinity. Since an infinite number of firms per industry is unrealistic, the assumption is made that as long as there are "many" firms, each acts "as if" there were an infinite number and this produces perfect competition. Firms, therefore, act as if they are price takers when in fact they are not.<sup>1</sup>

In this paper, we initially describe a partial equilibrium model—loosely called Cournotian, after Cournot—in which there is a positive relationship between the degree of competition and the number of firms in an industry. We then proceed to show that this relationship disappears in a general equilibrium model. In fact, the major result of the general equilibrium analysis is the following: under certain conditions, a general equilibrium with two or more noncolluding firms per industry is perfectly competitive.

## I. The Cournot Model: Firms Mind Their $p$ 's But Not Their $q$ 's

Assume that there is an industry demand

\* Professor and associate professor at the Graduate School of Business, University of Chicago, respectively. The paper has benefitted from the comments of Fischer Black, George Borts, Jacob Frenkel, John Gould, Michael Jensen, and Merton Miller. Financial support was provided by the National Science Foundation.

<sup>1</sup> A representative statement made by Armen Alchian and William Allen is the following:

We use the term "price takers' markets" to describe a class of markets where every supplier (and also every demander) provides so small a portion of the supply (or demand) that his output (or demand) has no significant effect on price; hence he "takes" the market price as if it were given by outside forces. [p. 106]

Similar and somewhat more detailed statements can be found in G. L. Bach, p. 370 and Paul A. Samuelson, p. 516.

curve,  $q = q(p)$ , relating the quantity of the good demanded,  $q$ , to the price of the good,  $p$ , and that  $q$  has an inverse function,  $p = f(q)$ , which gives the price the market is willing to pay for  $q$  units of the good. It is assumed that both  $q$  and  $f$  are everywhere differentiable, and that the demand curve is negatively sloped, that is,  $dq/dp < 0$ .

If we also assume for simplicity that the costs to any firm,  $i$ , are only related to that firm's output,  $q_i$ , the firm's total profits,  $\pi_i$ , are

$$(1) \quad \pi_i = q_i p - \phi(q_i) \\ = q_i f \left[ \sum_{j \neq i} q_j + q_i \right] - \phi(q_i),$$

where  $\phi(q_i)$  is the firm's cost function and  $\sum_{j \neq i} q_j$  is the total output of the rest of the firms in this industry. Profit maximizing output for firm  $i$  occurs at a point where marginal revenue equals marginal cost

$$(2) \quad p + q_i f'(q) \left( 1 + \frac{d \sum_{j \neq i} q_j}{dq_i} \right) = \phi'(q_i)$$

In terms of equation (2), the degree of competitiveness of this industry depends on

$$d \sum_{j \neq i} q_j / dq_i$$

the output changes by all other firms that firm  $i$  anticipates will come in response to a one unit change in its output. For example, suppose that for any firm  $i$ ,

$$d \sum_{j \neq i} q_j / dq_i = -1$$

so that a change in output by firm  $i$  is anticipated to be precisely offset by changes on the part of other firms. In this case, even though the number of firms is finite, the industry is perfectly competitive: The individual firm's output decision literally has no effect on the market price of the good,

and, from (2), profit maximizing output is where marginal cost is equal to price.

We eventually show that when there are two or more noncolluding firms in the industry, then under certain conditions a general equilibrium in fact implies the response function

$$d \sum_{j \neq i} q_j / dq_i = -1$$

so that the industry equilibrium is perfectly competitive. First, however, we consider the case, common in partial equilibrium treatments of firms and industry equilibrium, where other firms are not anticipated to change their output decisions in response to a change in the output decision of firm  $i$ —that is,

$$d \sum_{j \neq i} q_j / dq_i = 0$$

We find that the analysis of this case, which we call Cournotian, leads to relationships between the number of firms in an industry and the degree of competition like those described in the introduction.

When

$$d \sum_{j \neq i} q_j / dq_i = 0$$

equation (2) reduces to

$$(3) \quad p = \phi'(q_i) - q_i f'(q)$$

In terms of elasticities, we get

$$(4) \quad p = \phi'(q_i) + \frac{y_i}{\eta} p$$

where  $\eta = -(dq/dp)(p/q)$  and  $y_i = q_i/q$ . Equation (4) in turn simplifies to

$$(5) \quad p = \frac{\eta \phi'(q_i)}{(\eta - y_i)}$$

If all firms have the same technology and thus the same cost function, they make the same output decisions, and the share of any firm  $i$  in total output is  $y_i = 1/n$ , where  $n$  is the number of firms. Thus (5) can be rewritten

$$(6) \quad p = \frac{n\eta \phi'(q_i)}{n\eta - 1}$$

Here the relationship between the number of firms and the degree of competition in the industry is clear. Other things equal, the larger  $n$ , the closer is price to marginal cost. Alternatively, in the Cournot model the firm's perceived elasticity of demand is  $n\eta$ . In the limit, that is, as  $n \rightarrow \infty$ , the firm's perceived elasticity is infinite, price is equal to marginal cost, and the industry is perfectly competitive.

From a general equilibrium viewpoint, the Cournot partial equilibrium assumption (that individually firms behave as if their price-output decisions have no effect on the output of other firms) can be criticized in that it places firms in the position of assuming that other firms do not behave optimally. And criticisms from the general equilibrium viewpoint are critical when their consideration would overturn the results of the partial equilibrium analysis. Such is the case here. We show now that under certain conditions the Cournotian relationship between the number of firms and the degree of competition in an industry disappears in a general equilibrium analysis.<sup>2</sup>

<sup>2</sup> In fact, even in the Cournot partial equilibrium model the conclusion that "numbers count" can disappear when there is a perturbation in the conditions of equilibrium. For example, suppose that there is a change in the industry's price elasticity of demand. This could result from a shift in the demand curve or from a change in slope. In such cases it is in the interests of firms to alter their prices. Since all firms start from the same position and have the same perceptions, they all change price by the same amount. But in the Cournot model firms assume that other firms do not change their output in response to a change in price; that is, each firm perceives the slope of its demand curve to be  $dq/dp$ , the slope of the industry demand curve. Thus each firm expects to experience the entire change in the industry quantity demanded implied by the price change. In fact, however, the change in the quantity demanded is divided evenly among all the firms. If each firm persists in assuming that other firms did not change their outputs, then observation of the effects of the price change leads each to adjust its estimate of the slope of the industry demand curve. If the adjustment is complete, firms now consider the slope of the industry demand curve to be  $(1/n)dq/dp$  instead of  $dq/dp$ . Or equivalently, each perceives the elasticity of the industry demand curve to be  $(1/n)\eta$ . From equation (6), the new equilibrium price is

$$p = \frac{\eta \phi'(q_i)}{\eta - 1}$$

## II. The General Equilibrium Model

We are concerned with the usual static model in which tastes, endowments, technology, population, etc., are taken as given. The following conditions are also assumed to hold:

**C1:** Factors of production are infinitely divisible and fully and costlessly mobile across firms and industries. Moreover, either there are no factors that are used in only one industry, or such factors are owned by individuals and their services are sold to firms.

**C2:** Information about the returns earned by factors in different firms and industries is available costlessly to everybody, and factors act to maximize their returns.

**C3:** The same techniques of production are available to all firms in any given industry.

**C4:** The demand curve for the good produced by any industry is downward sloping with nonzero elasticity at all levels of output.

**C5:** At any given level of output, the curve showing the minimum marginal cost of industry output as a function of industry output has greater slope than the industry demand curve.

---

Comparing the above equation with equation (6), it is easy to see that the new price is precisely the price a pure monopolist would charge. And this result, which comes about without collusion, is independent of the actual number of firms in the industry.

Moreover, the result is much more general than the example of a shift in the industry demand curve. The Cournot equilibrium of equation (6) degenerates into the equivalent of a purely monopolistic industry equilibrium (irrespective of the number of firms in the industry) whenever (i) there is a shock that affects all firms in the industry in precisely the same way—for example, a common shift in cost curves—but (ii) in adjusting to the shock each firm persists in the Cournot assumption that other firms do not change their output decisions.

Finally, within the context of the Cournot partial equilibrium model, it is also possible to construct cases where there are shocks that affect firms differently, and where the result is that the industry goes from the Cournot equilibrium to a perfectly competitive equilibrium, irrespective of the number of firms in the industry. But rather than take this partial equilibrium

**C6:** Conditions of production are such that the rest of an industry can always use precisely the incremental quantities of factors of production demanded or released by a firm to offset precisely output changes by that firm.

Conditions **C1** to **C3** are standard. Conditions **C4** and **C5** are meant to rule out—and thus allow us to ignore—the types of “unusual” demand or cost conditions that could lead to problems with the stability of an industry equilibrium—problems that would only serve to confuse our arguments about the effects of the number of firms.<sup>3,4</sup> Finally, condition **C6** has two major implications. First, strictly speaking, production functions must be characterized by constant returns to scale. Second, offsetting output changes by the rest of an industry in response to an output change by an individual firm must also precisely offset any externalities associated with the firm’s new decision.

Given the maximizing behavior and costless mobility of factors and technologies assumed in conditions **C1–C3**, under conditions **C1–C6** a general equilibrium implies that returns to any factor are equal across all firms and industries.<sup>5</sup> We wish to show,

---

path to perfect competition, we prefer to concentrate on the general equilibrium arguments that follow.

<sup>3</sup> If the industry is perfectly competitive, then **C5** implies that at any given level of output, the industry supply curve has greater slope than the industry demand curve, so that equilibrium is unique. But **C1–C6** are meant to be a minimum set of conditions that imply perfect competition, which in turn implies that the minimum marginal cost of industry output as a function of industry output is also the industry’s supply curve. Thus, since the concern is with the minimum conditions that imply perfect competition, **C5** is not stated directly in terms of supply curves.

<sup>4</sup> It is well to note that conditions **C1–C5** are consistent with the Cournot model and are also commonly assumed in less specific analyses, like those referenced in the introduction, in which competition in an industry is assumed to depend on the existence of many firms, each producing only a small fraction of the industry output.

<sup>5</sup> Thus if there are monopoly rents in the returns to any factor, then these monopoly rents must be equal across all firms and industries. The existence of such rents has no effect on the analysis.

however, that if there are two or more non-colluding firms in each industry, such a general equilibrium is also perfectly competitive in the sense that individual firms literally face perfectly elastic demand curves for their outputs. That is, when conditions C1-C6 hold, with two or more firms per industry the output decisions of a firm literally have no effect on prices.

To establish this result, suppose that, for whatever reason (perhaps simply to test the slope of its demand curve), a firm decides to disturb the general equilibrium by increasing its output. Without loss of generality, assume that there are no associated reactions from firms in other industries. Given the demand condition C4, if other firms in the changing industry initially did not lower their output, the per-unit price of the good produced by this industry would fall relative to prices of other goods. Given the industry cost condition C5, returns at least to some factors would then be lower in this industry than in others. Thus movement back to general equilibrium—that is, equality of returns across firms and industries—implies that other firms in the industry contract their output in response to an expansion by an individual firm.

Moreover, given the demand and cost conditions C4 and C5, and given the constant returns and offsetting externalities implied by condition C6, if other firms in the industry contract their output by an amount less than the individual firm expands, factor returns are still too low in this industry vis-à-vis others. On the other hand, given conditions C4 to C6 factor returns in the industry are too high vis-à-vis other industries if other firms in the industry contract their output by an amount more than the individual firm expands. Thus given the maximizing behavior and costless mobility of factors assumed in conditions C1-C3, a return to the equality of returns implied by a general equilibrium requires that other firms in its industry respond by precisely offsetting any output increase by an individual firm.

Since analogous reasoning applies to output decreases, we can conclude that, given

conditions C1 to C6, the ultimate general equilibrium response by other firms in an industry to a change in output by an individual firm is to offset precisely that output change; that is, in terms of (2),

$$d \sum_{j \neq i} q_j / dq_i = -1$$

so that price per unit of output is unaffected by the output decisions of individual firms. In short, as long as there are two or more noncolluding firms in each industry and conditions C1-C6 hold, then a general equilibrium implies that each firm is literally perfectly competitive in the sense that it faces a horizontal demand curve for its output.<sup>6</sup>

It is helpful to reexamine the roles of conditions C1-C6 in this result. First, the frictionless mobility and maximizing behavior of factors assumed in conditions C1-C3 are instrumental in the conclusion that in a general equilibrium factors are distributed across firms and industries in such a way that the returns to any factor are equal across all firms and industries. Then the "regularity" or "stability" conditions on demand and cost functions assumed in conditions C4 and C5 ensure that this general equilibrium principle of equal returns implies that the rest of an industry acts to offset the output changes of an individual firm. Finally, the constant returns and offsetting externalities implied by condition C6 then guarantee that the general equilibrium response of the industry is to offset precisely any output changes of a firm, which in turn, of course, implies that the output decisions of a firm literally have no effect on prices.<sup>7</sup>

We emphasize that this analysis does not depend on any specific adjustment process by which other firms respond to output

<sup>6</sup> Note that this analysis *implies* that each firm is perfectly competitive, that is, faces horizontal supply curves in all factor markets, even though the industry as a whole may face upward sloping supply curves for factors.

<sup>7</sup> The offsetting output changes may come either from existing firms, including the possibility of exit, or from entry of new firms.

changes of an individual firm. Any adjustment process that converges to a general equilibrium—that is, equality of returns to any factor across all firms and industries—will do. And given conditions C1–C6, the general equilibrium is one in which individual firms are price takers.

Perhaps most objectionable in the conditions C1–C6 is the implication of C6 that production functions are characterized by constant returns to scale. With constant returns, though the output of the industry in a general equilibrium is determinate, the output or size of any individual firm is not. But except for some annoying “discontinuities,” our analysis can also hold when there are nonconstant returns to scale—more specifically, when there is a finite optimal firm or plant size. Then we can only assume that condition C6 holds for output changes by individual firms that are multiples of the optimal firm or plant size. For such output changes, it is again true that price per unit of output is unaffected by the production decisions of individual firms. In other cases the general equilibrium response by other firms to a change in output by an individual firm may be to offset only partially the firm’s output change, though the ultimate response must be sufficient to make it unprofitable for

another whole firm or plant to enter or leave the industry.<sup>8</sup>

### III. Summary

But even with the annoying discontinuities that arise when there is an optimal firm or plant size, the important general result still holds: When there are at least two non-colluding firms in an industry, there is no clear-cut relationship between the number of firms and the degree of competition. The absence of perfect competition must arise from violation of one or more of the conditions C1–C6—that is, it must arise from such things as indivisibilities, factor immobility, nonmaximizing behavior by factors, monopolistic access by individual firms to production techniques, and lack of information concerning the returns earned by given factors in different uses.

### REFERENCES

- A. A. Alchian and W. R. Allen, *University Economics*, Belmont 1967.
- G. L. Bach, *Economics*, Englewood Cliffs 1963.
- P. A. Samuelson, *Economics*, New York 1961.

<sup>8</sup> One is, of course, hard pressed to reconcile the assumption of infinitely divisible factors with the existence of a finite optimal firm or plant size. But we do not wish to get into that issue here.

# Soviet Postwar Economic Growth and Capital-Labor Substitution: Comment

By EARL R. BRUBAKER\*

In his article in this *Review*, Martin Weitzman carries out a series of well-constructed and useful statistical experiments assessing the consistency of a considerable portion of the record<sup>1</sup> on Soviet growth with several familiar a priori specifications of an aggregate production function with technical progress. On the criterion of minimum error sum of squares, Weitzman finds that the CES, Hicks-neutral specification appears to be as good an approximation to "reality" as any of the others. He finds this version especially useful for explaining the much heralded deceleration in growth of output, suggesting that aside from a geometric time trend the postwar Soviet growth record is adequately accounted for by a constant elasticity of substitution (CES) production function with elasticity of substitution significantly less than one. Thus he is making a rather novel interpretation of the statistics on Soviet growth. Many investigators had concluded that the slowdown in growth of Soviet output may be attributed to a declining residual. This conclusion was based, however, on procedures which often involved an essentially arbitrary assumption about the value of the elasticity of factor substitution ( $\sigma$ ), namely, that its value is one. By estimating econometrically the value for  $\sigma$ , Weitzman appears to remove an element of arbitrariness and to provide support for an explanation of the deceleration in the growth of Soviet output that emphasizes diminishing returns to capital accumulation rather than a declining residual. Unfortunately, in the light of additional pertinent evidence the

factor marginal productivities implied by the parameter estimates appear to be virtually incredible. Even if they were acceptable as a gross approximation to reality, Weitzman's data and estimated parameters are not consistent with one of his major conclusions, namely that diminishing returns are the principal explanation for the post-1965 slowdown in the rate of growth of industrial output. The purpose of this communication is to elaborate these two points.

The estimated marginal productivities of capital reflect an apparently far greater efficiency of investment decision making than appears possible in the light of the consistent theme of a substantial body of literature in the East and West.<sup>2</sup> For much of the period in question there was no interest charge to users of capital, and amortization charges were generally too small to affect investment calculations significantly. Lack of an explicit interest rate resulted in chronic understatement of real costs of capital, encouraged a bias toward irrationally capital-intensive project choices, hoarding, and generally profligate use of capital. It was not until the late 1950's that the authorities granted approval for a systematic procedure to aid in choice among projects. Even in the mid-1960's T. S. Khachaturov complained in *Pravda* that the methods economists had developed to measure the effectiveness of investment were being ignored. Also in the mid-1960's, the Soviet leadership apparently felt the situation so desperate as to require the ideologically execrable step of instituting net capital charges. Under the circumstances, Weitzman's estimates implying that the average gross rate of return in Soviet industry amounted to something on the order of 20 to 40 percent during the 1950's and

\* Associate professor, University of Wisconsin. Preparation of this comment was facilitated by a discussion in Professor A. Bergson's Seminar on Comparative Economics at Harvard University.

<sup>1</sup> The analysis is limited to an aggregation of industry, construction, transportation, communications, and distribution in one case and to industry alone in another.

<sup>2</sup> See Abram Bergson and Alec Nove for useful summaries and abundant references.

1960's are remarkable, to say the least. One can, of course, only speculate as to the rates of return that might have been achieved had the sources of blatant inefficiency, so troublesome for Soviet economists and authorities alike, been absent. Alternatively one may find that the econometrically estimated capital marginal productivities are far too high to be taken very seriously. Weitzman himself seems amenable to the latter approach, since on page 691 he finds that in 1960 a net rate of return of 10 percent, rather than the approximately 25 percent implied by his econometric estimates, seems "about right" in the light of U.S. experience.

To fully appreciate the unusual implications of Weitzman's estimates for relative marginal productivities of capital and labor in various nonagricultural sectors, it is helpful to examine his data on "imputed factor shares." Note first, however, that the latter are not based on the assumption that factors actually are paid their marginal products.<sup>3</sup> Imputed factor shares are, rather, those that would be generated by paying observed inputs their marginal products as estimated statistically. According to data in Weitzman's Table 3, in 1950 the imputed share for labor in Soviet industry amounted to 14 percent of the total as compared with 86 percent for capital. With the elasticity of substitution at only .4 and the growth of the capital to labor ratio quite high, by 1969 the labor and capital shares would have become 59 and 44 percent, respectively. The even more startling implication (not explicitly stated by Weitzman) for the more comprehensive sector, including industry, construction, etc., is an imputed labor share of approximately 9 percent in 1950 rising to

about 68 percent in 1966. Evidently as of 1950 in some nonagricultural, nonresidential sectors other than industry the imputed labor share would have been even lower than 9 percent.

We might note also that if the elasticity of substitution estimated for the 1950's and 1960's indeed reflects "reality," and if, as would seem plausible, the underlying factors determining its value operated in the 1930's as well, we would be faced with interpreting estimates of labor's share in 1928 even lower than the rather suspicious looking values for 1950. According to Richard Moorsteen and Raymond Powell, between 1928 and 1950 the tangible capital to labor ratio approximately doubled in the nonagricultural, nonresidential sector, and an elasticity of substitution much less than one would imply a substantial increase in labor's share in this period also. Apparently the marginal productivity of labor in much of the *nonagricultural* sector would have had to be virtually zero in 1928.

One certainly would feel more comfortable with all the above if it were possible to make reference to another economy which, with factor proportions similar to those of the USSR in this period and with observed factor shares presumably a reasonable approximation to imputations, showed results remotely similar to these. In Italy, for instance, according to Edward Denison, pp. 38-42, the share of labor amounted to 72-76 percent of national income from 1950 through 1962. The possible distortions of noncompetitive market power may come to mind, but as Bergson has remarked, to attribute to market power the difference between values as estimated for the USSR and those observed in Italy, one would have to adopt an untenable position regarding the relative bargaining strength of labor and tangible property.

Why should an econometric study by so masterful an econometrician yield such patently implausible results? Given the impressive evidence M. Ishaq Nadiri has collected showing current difficulties in attempts to estimate accurately the parameters of aggregate production functions for West-

<sup>3</sup> There is another sense in which the question of the relationship between marginal productivities and the observed remuneration to the factors is relevant. Do the highly aggregated output and capital measures have any meaning if the observed remunerations to the factors do not correspond to marginal productivities? This is an important question, but it is beyond the scope both of Weitzman's article and of this note. For the present any divergence of factor payments from their marginal products may be regarded as just another of the possible sources of error in the data.

ern economies, one can only ponder in awe how these difficulties may be compounded where the data have their underlying basis in the Soviet statistical and economic systems. Even where only industry is concerned, the known diversity in productive activities aggregated and the extent of structural shifts are sobering indeed. There is the assumption of constant returns to scale, convenient, but perhaps a substantial deviation from "truth." There is omission of inputs other than labor and fixed capital (Powell has found that material and inventories accounted for more than 20 percent of total industrial costs in 1950). Omission of educational capital could be important, but intangible capital other than educational, accumulated essentially through experience of the labor force, may be the factor which, neglected, could flaw the estimates most severely of all. The list of problems could be extended, but the essential point is clear. We have many good reasons for doubting that our econometric results can yield anything more than the grossest of approximations to the underlying "reality."

While bearing these reservations in mind, let us suppose for the sake of discussion that the CES function as estimated is after all an acceptable approximation. It would appear to follow that Weitzman's computed parameters and basic data are not consistent with his fundamental argument on page 685 that diminishing returns explain the post-1965 deceleration in growth of output.

In fact, a rather different interpretation appears warranted. To see why, it will be helpful to think in terms of equation (1) relating growth rates for output, the residual, and factor inputs.<sup>4</sup>

$$(1) \quad g_Y = g_A + [(1 - \eta_L)g_K + \eta_L g_L]$$

Since  $g_A$  has been constrained to a constant,

<sup>4</sup> For convenience, Weitzman's notation is adopted and repeated here.

$A$  = residual

$Y$  = output

$K$  = capital

$L$  = labor

$g_x$  = growth rate of subscripted variable

$\eta_x$  = share of subscripted factor

any decleration in  $g_y$  will have to occur as the result of a decline in the value of the quantity in brackets, hereafter referred to as  $z$ . The Weitzman interpretation holds that such a decline occurred essentially not because of declines in  $g_K$  or  $g_L$ , but because of diminishing returns, apparently manifested through an increase of the share of labor,  $\eta_L$ , since that is the only way left to reduce  $z$ . Thus rather than simply characterize generally the decline in  $z$  as due to diminishing returns, it might be useful to pinpoint the source of increase in  $\eta_L$  in terms of the various factors that influence  $d\eta_L$ . Clearly  $d\eta_L/dt$  is a function of  $\sigma$ ,  $g_{K/L}$ , and  $K/L$ . Now  $\sigma$  is assumed constant throughout 1950-69, and  $g_{K/L}$  in fact declines in the period 1964-69. Only the higher values for  $K/L$  would tend toward higher values for  $d\eta_L/dt$ . The data on industry alone in Weitzman's Table 3 (after adjustment so that  $\eta_L + \eta_K = 100$  each year) imply the opposite of the required effect, however, since they show that  $d\eta_L/dt$  averaged 2.4 percentage points per year in 1950-65 as opposed to only 1.8 percentage points per year in 1965-69. Thus it seems that diminishing returns would have contributed more heavily toward deceleration in growth during the early period.

If  $d\eta_L/dt$  was actually relatively lower in 1965-69, what was the source of the decline in  $z$ ? Consider the variation in  $g_K$  and  $g_L$  as presented in the following tabulation:

AVERAGE ANNUAL RELATIVE RATES OF GROWTH

Period	$g_K$	$g_L$
1950-65	11.4	2.5
1965-69	8.4	3.1
1950-60	11.5	2.3
1960-69	10.0	3.0

The three percentage point drop in  $g_K$  between 1950-65 and 1965-69 immediately arouses suspicions, especially since the values for  $\eta_K$  had fallen to little less than 50 percent by the latter period. The direct effect on  $z$  of a decline in  $g_K$  tends to be counter-balanced by an indirect effect in the form of a tendency towards less rapid gain in labor's weight, but in the given case this in-

direct effect would be about 0.1–0.2 percentage points per annum. Thus it seems that even if we accept Weitzman's preferred hypothesis, i.e., CES with  $g_A$  constant, we would have to explain the deceleration in growth of output after 1965 largely in terms of a declining  $g_K$  rather than in terms of an abnormally large (relative to 1950–65)  $d\eta_L/dt$ . Note that the data for the periods 1950–60 and 1960–69 permit a roughly similar argument.

Turning briefly to policy implications of the Weitzman analysis, it can be seen almost immediately that they will serve more as an indication of the utter implausibility of the estimated parameters than as a useful guide to action. Weitzman correctly states that granted: 1) the validity of his analysis and 2) its continued applicability in the future, "... it would appear that a strategy of strong capital accumulation must be considerably less successful for the present relatively mature Soviet economy..." (p. 685). But an explicit numerical statement is quite essential to grasp the bizarre implications of the estimated parameters. For the present this is clearest from a consideration of Weitzman's parameter estimates from the more reliable and more comprehensive data of Moorsteen and Powell. These parameters imply an imputed capital share falling from 91 percent in 1950 to about 32 percent in 1966. Granted that the model and numerical estimates continue to be appropriate, by now the Soviet planners should be looking forward to the consequences of  $\eta_K$  approaching zero. An important consequence is, of course, that they should be able to allow  $g_K$  to drop from the 9 percent a year

maintained in the 1960's to the value expected for  $g_L$ , say about 2 percent a year, with no significant decrease in  $g_Y$ . What a momentous opportunity! How regrettable it will be if, in the interests of science, we shall be unable to convince the Soviet leadership to try it.

In sum, even accepting the CES model as estimated, diminishing returns seem hardly the principal explanation of the deceleration of growth of output in Soviet industry in the latter 1960's. Considering the substantial grounds for doubting the accuracy of the estimated parameters, the case for diminishing returns as the dominant force virtually evaporates.

#### REFERENCES

- A. Bergson, *The Economics of Soviet Planning*, New Haven 1964.
- E. F. Denison, *Why Growth Rates Differ*, Washington 1967.
- R. Moorsteen and R. P. Powell, *The Soviet Capital Stock, 1928–1962*, Homewood 1966.
- M. I. Nadiri, "Some Approaches to the Theory and Measurement of Total Factor Productivity: A Survey," *J. Econ. Lit.*, Dec. 1970, 8, 1137–77.
- A. Nove, *The Soviet Economy*, 2d ed., New York 1969.
- R. P. Powell, "Industrial Production," in A. Bergson and S. Kuznets, eds., *Economic Trends in the Soviet Union*, Cambridge 1963.
- M. L. Weitzman, "Soviet Postwar Economic Growth and Capital-Labor Substitution," *Amer. Econ. Rev.*, Sept. 1970, 60, 676–92.
- Pravda*, March 15, 1966.

# Soviet Postwar Economic Growth and Capital-Labor Substitution: Comment

By MITCHELL KELLMAN AND LORENZO L. PEREZ\*

In a recent article in this *Review*, Martin Weitzman argued that the observable slowdown in the growth of output ( $g_y$ ) of the Soviet economy in the 1960's need not be associated with a fall in the growth of total factor productivity ( $g_a$ ), as is usually suggested, but rather can be better shown to be a manifestation of diminishing returns to capital. By directly estimating a Constant Elasticity of Substitution (CES) production function<sup>1</sup> for the two decades following World War II, he found an elasticity of substitution of capital for labor ( $\sigma$ ) significantly less than one. From this he concluded that the slowdown in the growth of that economy could largely be explained in terms of the diminishing returns to capital which resulted from the small substitutability between capital and labor and rapidly increasing overall capital deepening in the economy.

Weitzman concluded that "Instead of capital, labor and technical change will have to be increasingly relied upon as alternative sources of future economic growth" (p. 685); and [that due to demographic trends] "This rests the spotlight finally on technical change . . . the most appealing way of raising  $g_y$  is now to increase  $g_a$  . . . because  $g_L$  is more or less fixed . . ." (p. 686).

We should like to advance the proposition that the record of growth of the Soviet economy during the 1950's and 1960's (as presented in Weitzman's Table 1, p. 677) points to aspects of the underlying Soviet macro-production process other than the small elasticity of substitution as possibly the key culprits effecting the noted slowdown in  $g_y$ . Furthermore it is suggested that perhaps the "most appealing" way of raising

$g_y$  may after all be not through the overall productivity relationship  $A$  (or  $g_a$ ), but rather through the term slighted by Weitzman—the growth rate of the labor force  $g_L$ .

We fit the data in Weitzman's Table 1 to a maximum likelihood, non-linear regression program,<sup>2</sup> similar to that used by Weitzman. A more general model was employed which imposed neither a geometric time trend, nor unitary returns to scale on the data. The specification used was:

$$Y(K, L) = A[\delta K^{-\rho} + (1 - \delta)L^{-\rho}]^{-1/\rho}$$

where  $\rho = (1 - \sigma)/\sigma$  and  $0 < \rho < \infty$  indicates returns to scale for both factors.

This was run for two decades as a whole (equation (2)), and separately for each of the two component decades (equations (3) and (4)) as shown in Table 1.

TABLE 1

	$\hat{A}$	$\hat{\delta}$	$\hat{\sigma}$	$\hat{\rho}$	$R^2$
(1) Weitzman's findings <sup>a</sup>	.639 (.070)	.403 (.030)	con- straint to 1.		.9995
(2) Ours 1950-1969—	.641 (.112)	.536 (.146)	1.272 (.151)		.9987
(3) Ours 1950-1959—	.211 (.057)	.715 (.112)	— <sup>b</sup>	1.339 (.059)	.9999
(4) Ours 1960-—	.216 (.521)	.422 (.248)	— <sup>b</sup>	1.336 (.525)	.9970

Note: Numbers in parentheses are standard errors.

<sup>a</sup> Weitzman, p. 681.

<sup>b</sup> The standard errors of the respective decade's  $\rho$  were large so that the values of the parameters were not significant.

The findings in equation (2) substantiate Weitzman's finding of an elasticity of substitution less than one for the period as a whole. If indeed this was the main reason

\* Instructor at Lafayette College, Pennsylvania, and research associate at the University of Pennsylvania, respectively.

<sup>1</sup>  $Y(t) = A(t)[\delta K(t)^{-\rho} + (1 - \delta)L(t)^{-\rho}]^{1/\rho}$  where  $A(t) = \gamma e^{\lambda t}$  see Weitzman, equation (4), p. 680.

<sup>2</sup> See Bard. Program set to a CES specification with the kind assistance of Fred McElroy at Georgetown University.

for the slowdown of  $g_v$ , its effect must have been quite powerful, acting as it did in a time of rapid overall growth of an economy characterized by increasing returns to scale ( $\hat{\mu} > 1$ ).

However, another factor emerged when the regression was estimated separately for the 1950's and the 1960's, respectively.<sup>3</sup> The "capital-intensity" parameter  $\delta$  fell from .72 in the first decade, to .42 in the second.<sup>4</sup> Generally a fall in  $\delta$  represents a (nonneutral) change which is capital saving, in the sense that capital's marginal product falls relative to labor's, for each capital-labor ratio; and hence for a given marginal rate of substitution a larger labor-capital ratio is required. Conversely such a change can of course be considered to constitute a labor-using technological change.<sup>5</sup>

If an effective, rational process of induced innovation had been operating in the Soviet economy during this period, one would have expected the rapidly rising capital-labor ratio to be reflected in an increasingly capital-intensive technology. Such a development would have allowed that economy to accommodate (in a dynamic sense) the increasingly abundant capital per worker. However, our finding points exactly in the opposite direction. In the 1960's, the Soviet economy actually acquired a more labor-intensive nature than had been true of that economy in the 1950's. This means that, perversely, at the very time that labor shortages were developing in the Soviet economy, its technological structure was becoming more labor-using. A drop in the capital-intensity parameter at a time when capital is rising rapidly relative to labor has a strong negative output effect.<sup>6</sup>

Thus, the disturbing conclusion emerges that if past tendencies can be at all relied upon to tell us something of the near future, then encouragement of the residual element ( $g_a$ ) or "technical change" may after all not

be "the most appealing" or sensible way of increasing ( $g_v$ ). The direction of the technical change to date seems to have been counterproductive. The negative output effects of the perverse shift in the "intensity" of the technology, could have easily counterbalanced increases in the level of overall productivity. In any case, a look at the unchanged neutral scale efficiency parameter  $A$  in equations (3) and (4) indicates that  $g_a$  was effectively zero during this period. Thus, the overall effect of technological change upon output seems to have been negative.

However, we are not left solely with this pessimistic conclusion. Our regressions suggest that the Soviet economy is characterized by overall increasing returns to scale ( $\hat{\mu} > 1$ ). This was true for the overall sample periods as well as for each of the two subsamples (although the evidence for this tendency in the 1960's was perhaps weaker, as indicated by the larger standard error of  $\hat{\mu}$  in equation (4)). This is not an unexpected finding when we consider the continual large scale exploitation of newly developed raw materials and land that took place in this period. In his article, Weitzman documented a conscious official program resulting in the continued reduction in the average length of the work day and of the work week in the Soviet Union throughout this period (see p. 687). If this variable can still, within certain bounds, be considered an operational policy tool, it is clear that with increasing returns to scale in capital and labor, the potential returns to a retardation in this downward trend of labor man-hours are quite large—larger than would be indicated by a production function with returns to scale a priori constrained to one.

As a conclusion, the increase in the labor intensity of the overall Soviet economy leads to interesting speculation concerning the cost of the increasingly complex planning and administrative bureaucratic structure to growth efforts, and leads one to at least question the proposition that the ideas of Kantorovich and Liberman have to date affected the basic structure of the economy. This "labor-using" shift in the nature of the Soviet economy could have caused the ob-

<sup>3</sup> See the Appendix.

<sup>4</sup> Using a difference in means test, the respective values of  $\delta$  were found to be significantly different at the 88 percent level. See the Appendix.

<sup>5</sup> See Murray Brown (1966).

<sup>6</sup> See Brown (1966) p. 58.

served slowdown in output, with or without a small elasticity of substitution. The cost of allowing the rate of growth in the labor force to slow down (to the extent that it could have been politically controlled) was larger than one may have suspected due to potential increasing returns characterizing the Soviet economy.

Finally, it must be noted that the high standard errors in equation (4) for 1960-1969, suggest that the CES production function becomes a less reliable specification for the macro-Soviet economy as it evolves into a more modern and complex form (more like that of the United States). Thus caution is warranted in relying too heavily on this tool for analysis of more recent trends and for future projections of this economy.

#### APPENDIX

Methodologically, it may be argued that during a given technological era, the Diamond and McFadden Impossibility Theorem applies here, and negates the possibility of measuring the change in the bias (intensity) of the economy.<sup>7</sup> We evade the difficulty utilizing the rationale used by Brown and Cani,<sup>8</sup> namely that the two decades represent two distinct "technological epochs"; an assumption which allows for the requisite parameter identification.

In order to test the hypothesis that the two decades do in fact represent two distinct

technologies (and hence production functions), a Chow test was applied to the residual sum of squares for the whole period ( $H_1$ ), and for each of its component periods ( $J_1$  and  $K_1$ ).<sup>9</sup>

$$F(p, n + m - 2p)$$

$$= \frac{H_1 - J_1 - K_1}{J_1 + K_1} \frac{n + m - 2p}{p}$$

where  $p$  is the number of parameters,  $n$  is the number of observation in first period,  $m$  is the number of observations in the second period.

$$F(4, 12) = \frac{85.58 - .525 - 45.07}{45.07 + .525}$$

At the 90 percent significant level, an  $F = 2.48$  indicates that a structural change has occurred. The finding of  $F = 2.63$  therefore allows us to accept the hypothesis.

#### REFERENCES

- J. Bard, "Nonlinear Parameter Estimation and Programming," IBM, Dec., 1967.
- M. Brown, *The Theory and Empirical Analysis of Production*, Nat. Bur. Econ. Res., *Stud. Income and Wealth*, Vol. 31, New York 1967.
- , *On the Theory and Measurement of Technological Change*, Cambridge 1966.
- M. L. Weitzman, "Soviet Postwar Economic Growth and Capital-Labor Substitution," *Amer. Econ. Rev.*, Sept. 1970, 60, 676-92.

<sup>7</sup> See Brown (1967).

<sup>8</sup> See Brown (1967), p. 99.

<sup>9</sup> See Brown (1966), p. 115.

# Soviet Postwar Economic Growth and Capital-Labor Substitution: Reply

By MARTIN L. WEITZMAN\*

What I take exception to in Earl Brubaker's comments is not so much *what* he says as *how* he says it. If he were to have phrased things a little more moderately I might almost have agreed with some of the things he said. But the tone of finality about a subject which hardly lends itself to final judgements of that sort may give his remarks an undeserved impression of substance which ought to be removed.

Those parts of Brubaker's comments which touch upon difficulties of aggregation, inaccurate measurement of "true" inputs, etc., are really universal complaints about using aggregate data that we have heard many times before. I fully and unconditionally agree that they throw up serious obstacles to the measurement, interpretation, and application of aggregate production functions. I thought that this attitude was spelled out clearly enough in my paper. But what else can you do? If you want to get an overall picture you've got to use aggregate statistics and aggregative models, flawed as they may be in theory or practice. As Robert Solow has quipped, "It may be crooked, but it's the only wheel in town." Brubaker has not inhibited himself from being a heavy consumer of highly aggregated numbers in order to talk about and interpret overall Soviet performance. And who can fault him for that? But isn't it then a little unfair for him to make such an issue about this very point in criticizing someone else's work?

Then there is the old bugaboo about "Soviet Statistics." According to Brubaker "one can only ponder in awe how these difficulties (of production function estimation) may be compounded where the data have their underlying basis in the Soviet statistical and economic systems." This strikes

me as an unwarranted exaggeration. There *are* some added real problems associated with (today's) Soviet statistics, as I indicated in the Data Appendix to my paper. But there may also be areas of equal weight in which their statistics are more accurate and comprehensive than ours (not any worse would really be a more apt description). For example, in compiling capital stock estimates, as I indicated in my paper, the Soviet Central Statistical Institute has access to such data as the value of unfinished construction, the value of capital retirements, and value of capital stock as ascertained by periodic inventories. None of these are available for the United States. The lack of unemployment in the *USSR*, as I also pointed out, eliminates what is one of the major headaches for production function estimation in the West. I certainly don't want to be driven into claiming any more for the Soviet data I used than that to a first approximation they are no more unfit for direct production function estimation than *U.S.* data (this is *really* damning by faint praise!).

Brubaker doesn't like the fact that the imputed factor shares based on the parameter point estimates don't correspond closely enough to Western empirical factor shares. Now it's one order of magnitude to sin against economic theory (as I have) by postulating an aggregate production function and setting about measuring it. It's yet another level of sin to invoke a marginal productivity theory of distribution on one big production function and then identify factor shares with parameter values, especially for the *USSR*. What's more, there's nothing sacred about the derived parameter point estimates—I certainly wouldn't be willing to go to the stake for them. If my model has any relation to reality it has to be a very rough one. While the statistical tests performed seemed to in-

\* Associate professor of economics, Yale University.

dicate that  $\sigma < 1$ ,  $\sigma$  might be higher than its reported value, and  $\delta$  might be a little different. It seems to me that parameter values could be nudged around a little to make imputed factor shares look more like Italy's, if that's important. (Incidentally, the share of capital in Italian *manufacturing* is what we ought to be talking about, and that should be higher than it is for the Italian economy as a whole.) In statistical language, I believe that if all the regressions were rerun under the additional a priori specification that the empirical share of labor be between, say, .6 and .8, there would still be enough statistical power to force through the main conclusions, including  $\sigma < 1$ . The less-than-one elasticity of substitution for Soviet industry seems to be a hard result to escape unless you're willing to buy a pretty steep decrease in the residual's rate of growth.

I'm a little confused about where Brubaker finds the possibility of getting rate of return calculations from my work. Unlike the estimates of distribution shares, which are invariant to the units of measurement, rates of return would depend critically on how capital and output are being measured. In my work I expressed capital and output as artificial indexes with 1960 values arbitrarily normalized to 100. In order to measure meaningful rates of return one would in addition have to know the capital-output ratio in the usual units. This is a bit of a headache for a Soviet-type economy because Western-style data are not usually collected on the total net value of output (say for industry). If for the sake of argument an industry capital-output ratio of 2.5 is assumed along with 6 percent depreciation, the imputed rate of return in 1960 ought to be about 20 percent. I suppose this might be considered somewhat high, but the model is after all just a crude approximation and changing parameter values slightly could easily bring this figure down. And who knows, maybe it's the right value. Today about a 10 percent (net) rate is used for project evaluation in the *USSR*. Kantorovich and others have argued that this is much too low and it ought to be of the order of 15 percent or higher to more

accurately reflect real investment opportunities. In the early 1950's the rate of return should have been higher still. Brubaker sees some kind of a contradiction between this and the fact that there were no official interest charges in the early fifties. Of course capital was rationed in those years and its implicit rate of return must have been very high indeed, even if its nominal rate was zero.

I wish that Brubaker would have trained his guns on my "preferred" industry results and left the more comprehensive amalgam of industry, construction, transportation and distribution out of the action instead of pretty much the other way round. As I unambiguously stated I consider the data for this latter economic colossus more unreliable (for a variety of reasons) and only used it as an example to show that you get pretty much the same results if you work with data of "Western" origin instead of with data more directly derived from official Soviet statistics. It is this same desire to put up a front of respectability that motivated me to use 10 percent as a net rate of return in forming value-added weights for an index of industrial production. As I clearly stated, nobody knows too much about what rates of return ought to be anyway, and a rate twice that chosen would not have affected any of the results.

As far as which is "more important" in explaining the Soviet postwar industrial growth record—decreasing returns or the rate of growth of capital—this is a little like asking what is more important for life—sunshine or water. Both are important! I don't really understand the numerical demonstrations aimed at showing that the decline in  $g_k$  is more important than the decline in  $\eta_k$  in explaining the decelerated growth of output. There is a complicated tug of war going on here and it is difficult or impossible to unscramble long-term separate effects in a simple and meaningful way. If Brubaker wants to show that decreasing returns doesn't play much of a role, or plays less of a role after 1960, he should find a parameter to measure decreasing returns (maybe the

elasticity of substitution, but maybe something else) set up a statistical model and test away. The unscrambling of effects is one of the main things statistics is for! Those arithmetic exercises just beg the question.

In conclusion, the basic idea I was trying to foster in my exercise is that roughly speaking, there may have been some difficulties in substituting capital for labor in postwar Soviet industry. In explaining the Soviet postwar growth record, this friction can be an alternative explanation to the standard severe decrease in the rate of growth of total factor productivity. I attempted to demonstrate this proposition by quantifying it and then testing it using statistics. Without doubt there are lots of problems associated with data, model specification, statistical procedures, parameter estimates and interpretations. But the rough coherency and crude reasonableness of the results plus the overwhelming statistical power given to the proposition that the elasticity of substitution is less than one, seem to me to make a pretty good case.

The Kellman-Perez comment, in my opinion, is based on a general misuse of statistics. There isn't any good statistical reason for arguing that Soviet industry is characterized by overall increasing returns to scale. In fact, this may or may not be the case, but it emerges from the Kellman-Perez regressions only because they leave out a specification for year-to-year technical

change. Both are possible explanations of residual growth and their effects, as is well known, are usually indistinguishable.

Arguing that technical change is zero because  $A$  in both periods is about the same at .21 is ridiculous given the linearized standard error of .52. This comment is true even without addressing the more subtle issue of whether conclusions about (assumed) epochal technical change jumps can be used to shed light on the ordinary year-to-year garden variety of technological progress.

The Kellman-Perez two epoch no-technical-change-nonunitary economies of scale approach is of course not "more general" than mine, as the authors argue, but merely "different." This alone invalidates direct statistical comparisons, strictly speaking.

The Chow test which is employed measures difference in overall structure and cannot be used to verify, as is claimed, that  $\delta$  has changed. In fact, the latter conclusion seems downright unlikely given the relatively high linearized standard errors for  $\delta$ , but in any case, it is not demonstrated.

Incidentally, something seems to be awfully fishy about the numbers used in that Chow test. Can the error sum of squares in one subperiod be almost a hundred times greater than in the other?

In view of how weak Kellman and Perez's statistics are, it doesn't seem worthwhile to discuss the rather elaborate conclusions which are drawn from their use.

# A Model of Soviet-Type Economic Planning: Comment

By J. M. MONTIAS\*

Michael Manove's recent contribution in this *Review* is fundamental in at least two respects. For the first time he has shown precisely how the formulation of each year's plan could be facilitated by taking advantage of data derived in the course of constructing the plans of previous years; and he has demonstrated the relationship between the changes in supply and demand generated in the present year to the imbalances of previous years.

Although the logic of Manove's models is unassailable, I do have some reservations on the way he interprets, or seems to interpret, what is going on in some of these iterative processes he describes. It clearly emerges from his paper, in particular, that a procedure that calls for more iterations will necessarily produce absolutely smaller imbalances in the long run, for given yearly final demands and increments in output. In other words, the absolute value of the elements of the vector of supply-demand imbalances in year  $T$  will be smaller under the centralized procedure that includes "retrospective iterations" (Manove's equation 21), than under the procedure that does not (equation 20), provided that the input-output matrix  $A$  is productive. In symbols,  $|E_T| < |E_T'|$ , where

$$(1) \quad E_T = AL_T + A^3L_{T-1} + A^5L_{T-2} + \dots + A^{(2T-1)}L_1$$

$$(2) \quad E_T' = AL_T + A^2L_{T-1} + A^3L_{T-2} + \dots + A^TL_1$$

In the above expressions,  $L_t = Q_t - (AQ_t + \Delta D_t)$  ( $t=1, \dots, T$ ), where  $Q_t$  is a vector of output increments (or decrements) and  $\Delta D_t (=D_t - D_{t-1})$  is a vector of additions to final demand in period  $t$ .

The sign of the above inequality depends critically on the signs of the successive  $L_t$ . It will only hold in the general case if every  $L_t$  is composed exclusively of nonpositive elements. The following example will show that the result will not be valid in case certain reasonable values are postulated for the increments in supply and demand that will cause the vectors  $L_t$  to alternate in sign. To simplify the presentation, I have based my counter-example on a one-sector economy, but this extreme aggregation in no way affects the demonstration.

In this three-period process, the initial values of demand  $D_0$  and of the output increment  $Q_0$  are zero; in the next period,  $D_1=100$  and  $Q_1=100$ ; in the last period,  $Q_2=50$ ,  $D_2=100$ . The input-output matrix  $A$  collapses to a single element  $a$ , equal to 0.5.

We shall first follow the procedure involving central planning ( $T=1$ ) but no retrospective iterations ( $\Omega=0$ ). In the equations below,  $\hat{y}_t$  denotes the tentative target of year  $t$ ,  $y_t$ , the actual output of period  $t$ , and  $x_t$ , the total demand for that period ( $t=0, 1, 2$ ).

$$(3) \quad \hat{y}_1 = y_0 + Q_1 = 0 + 100 = 100$$

$$(4) \quad y_1 = \hat{x}_1 = a\hat{y}_1 + D_1 = 50 + 100 = 150$$

$$(5) \quad x_1 = ay_1 + D_1 = 75 + 100 = 175$$

$$(6) \quad E_1' = y_1 - x_1 = -25$$

There is thus a deficit, or excess of demand over supply, in period 1.

$$(7) \quad \hat{y}_2 = y_1 + Q_2 = 150 + 50 = 200$$

$$(8) \quad y_2 = \hat{x}_2 = a\hat{y}_2 + D_2 = 200$$

$$(9) \quad x_2 = ay_2 + D_2 = 200$$

$$(10) \quad E_2' = y_2 - x_2 = 0$$

Supply and demand are equal in period 2.

\* Yale University.

Substituting the above values of  $Q_t$ ,  $D_t$ , and  $a$  in equation (20) of Manove's paper, we obtain the same result:

$$(11) \quad E'_2 = .5[50 - (25 + 0)] \\ + .25[100 - (50 + 100)] = 0$$

We note that  $L_2$  (in the first set of brackets) is positive and  $L_1$  (in the second set of brackets) is negative and absolutely larger than  $L_1$ . However,  $aL_2$  cancels out with  $a^2L_1$ .

Let us now carry out the same exercise for  $\Gamma=1$  and  $\Omega=1$ , that is, for the case where retrospective iteration takes place in addition to central planning.

The results for period 0 and period 1 are identical:  $y_1=150$ ,  $x_1=175$ , and  $E_1=-25$ .

The difference comes in period 2 when  $x_1$ , the total demand emerging in period 1, replaces  $y_1$  in the equation for calculating the tentative output target of period 2.

$$(12) \quad y_2 = x_1 + Q_2 = 175 + 50 = 225$$

$$(13) \quad y_2 = \hat{x}_2 = .5(225) + 100 = 212.5$$

$$(14) \quad x_2 = .5(212.5) + 100 = 206.25$$

$$(15) \quad E_2 = y_2 - x_2 = 6.25$$

Substituting our basic parameters in equation (21), we also obtain:

$$(16) \quad E_2 = .5(50 - 25) + (.5)^2(-50) = 6.25$$

This time the weight applied to  $L_1$  (in the first set of parentheses) is too small to offset the positive effect of  $L_2$  (in the second set). Hence supply turns out to exceed demand, and  $|E_2| > |E'_2|$ .

But there is a more serious hitch in Manove's retrospective procedure. I am prepared to argue that, when capacity limitations are taken into account, the adoption of this procedure would actually increase planners' tension and would therefore be counterproductive.

Consider the following equations, which are derived by elementary substitution, from Manove's equations (4), (5), and (10).

When both  $\Gamma$  and  $\Omega=I$ , that is, where "central planning" and "retrospective iteration" both occur for all goods:

$$(17) \quad Y_t = A(X_{t-1} + Q_t) + D_t$$

When  $\Omega=0$  and only central planning takes place, we have:

$$(18) \quad Y_t = A(Y_{t-1} + Q_t) + D_t$$

Manove assumes that each period's planned outputs are precisely realized. They therefore satisfy any existing constraints on capacity. It may also be surmised, especially if the plans are taut, that in at least some sectors with a well-defined capacity, actual output will be precisely equal to capacity. Furthermore, the planned increases in output are likely to be identical with the additions in capacity accruing in period  $t$ .<sup>1</sup> Hence the vector  $(Y_{t-1} + Q_t)$  should be a fairly good approximation to the constraints on capacity to which the economy will be subject in period  $t$ , at least for those sectors where capacity was fully utilized in period  $t-1$ .

Suppose, to simplify matters, that the planners were able to pick a bill of final demand  $D_t$  such that each element of  $Y_t$  turned out to be smaller than or equal to each element of  $Y_{t-1} + Q_t$  or, in other words, that each output target was smaller or equal to the unique capacity constraint to which it was subject.

If  $Y_t$  is substituted for  $(Y_{t-1} + Q_t)$  in equation (18) above, the above assumption implies:

$$(19) \quad Y_t - AY_t - D_t \geq 0$$

But the expression on the left-hand side of this inequality equals  $Y_t - X_t$ , the surplus-supply vector. It follows that all the elements of this vector must be nonnegative, the zero elements corresponding to outputs that were precisely equal to their capacity limits. If it happened that demand in all previous periods had been set with due regard to capacity limitations, and all the imbalances  $E_t$  were of positive sign, additional internal iterations would reduce the surpluses and eventually bring about a perfect balance without any of the outputs transgressing their limits. In the absence of further iterations, the surpluses would emerge in the course of the execution

<sup>1</sup> Some elements of  $Q_t$  might conceivably be smaller than the planned capacity increases, although it is difficult to see how the planners could figure out in advance of the balancing process that not all the increments in capacity should be utilized.

of the plan as unanticipated increases in inventories. The planners could then assign these increments to final demand. The only shortcoming inherent in this truncation of the iterative process is that the resulting composition of net outputs might be less desirable than that which would have resulted from running the additional iterations.

We now observe from equation (17) above that if retrospective iteration had taken place with the same demand vector  $D_t$  as in equation (18) and if all the elements of the surplus-supply vector had been negative in period  $t-1$ , so that  $X_{t-1} > Y_{t-1}$ , some elements in  $Y_t$  would have exceeded the capacity limits set by the vector  $Y_{t-1} + Q_t$ . In this eventuality, Manove's scheme would break down, inasmuch as he assumes that every set of outputs  $Y_t$  is feasible.<sup>2</sup>

On the other hand, if a new demand vector  $D'_t$  had been set in such a way that the vector of outputs in equation (17) above did satisfy the capacity targets imposed by  $(Y_{t-1} + Q_t)$ , the net effect of introducing  $X_{t-1}$  instead of  $Y_{t-1}$  in the equation defining  $Y_t$  would have been to reduce the demand vector below its potential. It is improbable that the increments in this vector that would be made possible during the year  $t$  as a result of the allocation of surplus output would result in a total bill of net outputs as desirable as the original  $D_t$ .

The procedure I suggested for attaining consistency in the centrally planned balances on the basis of equation (18)—fitting the demand vector to the capacity constraints—of course does not require any explicit reference to the outputs of past periods. A still more efficient procedure, and one which I believe has some counterpart in planning practice, is for the planners to fix targets of net output only for commodities that are not capacity limited. The net outputs of the capacity limited sectors are then residual. As long as there are no crucial interdependencies among the sectors that are not capacity-

limited, this procedure requires at most one or two internal iterations to achieve a reasonable degree of consistency. I have argued elsewhere that industries producing finished goods are less likely to be capacity-limited sectors than those producing raw materials and semifabricates. Since interdependencies among sectors producing finished products are weak or inexistent, the imbalances due to incomplete iteration are likely to be small when these conditions are met.

To summarize my main points so far: (1) Retrospective iteration, which calls for the setting of tentative targets on the basis of the total *demand* generated in the previous period plus increments to output set for the current period, may lead to larger imbalances than the procedure relying on the actual *outputs* of the previous period if positive and negative vectors of imbalances alternate; (2) when capacity limits are taken into account, retrospective iteration is likely either to generate gross outputs in excess of these limits or to generate an inferior bill of final demands in the course of the plan execution.

If these arguments are valid, where does this leave Manove's time-phased iterative process? For centrally planned commodities, the procedure omitting "retrospective iteration" (summarized in equation (18) above) may still help to explain why imbalances are not as serious as they might be despite the truncation of the internal iterative process. But I am especially intrigued by the possibilities of his model regarding locally produced commodities. For here there can be no conscious coordination of inputs with outputs or tailoring of the bill of net outputs to capacity limitations, since no single agency is in possession of all the necessary coefficients. What disturbs me though is that the iterative process summarized in Manove's equation (18) hinges on retrospective iteration. Without it, convergence would not take place at all.<sup>3</sup> But it is again evident that if the

<sup>3</sup> When  $\Gamma=0$  and  $\Omega=I$ , the three equations describing the process in period  $t$  are:

$$\hat{Y}_t = X_{t-1} + Q_t$$

$$Y_t = \hat{Y}_t$$

$$X_t = AY_t + D_t$$

(over)

<sup>2</sup> Some transgression of capacity limits must occur under these assumptions even if linear combinations of outputs, rather than individual outputs, are constrained by capacity limits, provided one or more of these constraints are binding in the procedure represented by equation (17) above.

vector of supply-demand imbalances was negative in period  $t-1$ , retrospective iteration in period  $t$  must lead to the transgression of one or more capacity constraints, provided that the output of an least one good was constrained by capacity in the period  $t-1$ .<sup>4</sup>

We are left with built-in tension in the local plans and no explicit way of curtailing output in the course of plan execution to satisfy the constraints.

The discussion so far has adhered faithfully to Manove's model of economic life in Marozhenoe, at least as I understand it. Leaving aside the problems the model presents, we are also entitled to ask how the

---

It is obvious that if  $\hat{Y}_t$  were equal to  $Y_{t-1} + Q_t$ , instead of  $X_{t-1} + Q_t$ , there would be no convergence process at work.

<sup>4</sup> If there was at least one good in deficit supply in period  $t-1$ :

$$X_{t-1} \geq Y_{t-1},$$

with a positive difference for at least one good  $i$ . Hence:  $X_{t-1} + Q_t \geq Y_{t-1} + Q_t$  where the left-hand side of the inequality represents output in period  $t$  and the right-hand side, the capacity constraints of this period, on the assumption that at least one element of the output vector  $Y_{t-1}$  was capacity limited in period  $t-1$ .

Marozhenoe model precisely mirrors planning practice in Soviet-type economies. Unfortunately, I doubt whether we have enough knowledge of planners' detailed procedures to come to any definite conclusions on this point. While I have seen no evidence, for example, that the previous year's total demand—whether it has been satisfied or not—is added to anticipated increments in output, neither can I deny that this retrospective iteration does occur, particularly in incompletely coordinated planning at lower levels. At the very least, we are to be thankful to Manove for inducing us to look at the planning process in an original way and for prompting us to ask concrete questions about planners' procedures we would not have thought of, had he not demonstrated their critical importance.

#### REFERENCES

- M. Manove, "A Model of Soviet-Type Economic Planning," *Amer. Econ. Rev.*, June 1971, 61, 390-406.  
 J. M. Montias, "On the Consistency and Efficiency of Central Plans," *Rev. Econ. Stud.*, Oct. 1962, 29, 280-93.

# A Model of Soviet-Type Economic Planning: Reply

By MICHAEL MANOVE\*

J. M. Montias is primarily concerned with the effectiveness of the retrospective iteration, the only completely decentralized planning procedure described in my model. He has two main points.

First, he shows that when production in every sector is centrally planned, the addition of a retrospective iteration to the planning procedure may actually increase supply-demand imbalances.

Secondly, he argues that given the existence of capacity limitations, the "retrospective iteration is likely either to generate gross outputs in excess of these limits or to generate an inferior bill of final demands in the course of the plan execution."

Both of Montias's points have very important implications for the interpretation of the model, and I would like to comment on them.

## I. The Retrospective Iteration and Supply-Demand Imbalances

Montias maintains that a retrospective iteration may have the effect of increasing the magnitude of the supply-demand imbalances if the  $L_t$  are not all of the same sign.<sup>1</sup> He provides an example to illustrate his point.<sup>2</sup> In order to evaluate the significance of his example, it is important to ascertain whether the example describes a situation that is likely to occur with some frequency,

\* Associate professor of economics, University of Michigan. I wish to acknowledge financial support from the Comparative Economics Program of the University of Michigan.

<sup>1</sup> The vector  $L_t$  represents the imbalance between the exogenous increment to gross output in year  $t$  and the exogenous increment to total demand in year  $t$ . See my article, p. 399.

<sup>2</sup> If  $A$  is a productive technology matrix, then  $A^t$  goes to 0 as  $t$  gets large, but the convergence process is not necessarily monotonic. Consequently, even if all of the elements of all of the  $L_t$ 's have the same sign, it is possible that for some sectors, at least,  $|E_T| > |E_T'|$  as Montias defines them. However, trials with Vladimir Trembl's 38-sector version of the Soviet technology matrix for 1959 suggests that this is not likely to occur with any frequency.

or whether it is merely a cleverly contrived freak.

Let  $E_2$  denote the supply-demand imbalance in period 2 resulting from the use of a procedure which includes central planning ( $\Gamma = I$ ) and a retrospective iteration ( $\Omega = I$ ), and let  $E_2'$  denote the imbalance generated by the same procedure without a retrospective iteration ( $\Gamma = I$ ,  $\Omega = 0$ ).

Invoking the simplifications that Montias employs, we have from equations (20) and (21) in the original paper that

$$E_2 = aL_2 + a^3L_1$$

and

$$E_2' = aL_2 + a^2L_1$$

where  $a$  is the one-sector version of the technology matrix  $A$ . We set  $a$  equal to .5 as Montias does (a crude approximation of the characteristic root of the Soviet 1959  $A$ -matrix), but instead of specifying specific values for  $L_1$  and  $L_2$ , we assume that  $L_1$  and  $L_2$  are random variables uniformly distributed between  $-L$  and  $+L$ , where  $L$  is some limit on the magnitude of the  $L_t$ 's. Given that  $L_1$  and  $L_2$  are independent the probability that  $|E_2'| < |E_2|$  is 41 percent.<sup>3</sup> Thus (much to my surprise, I confess), the Montias example turns out not to be a freak: given the above assumptions we can expect the retrospective iteration to be counterproductive in almost half the occasions it is used.

But our computation paints a gloomier picture of the usefulness of the retrospective iteration than is justified. The retrospective iteration can be counterproductive only when the  $L_t$  differ in sign, but in this situation the supply-demand imbalance tends to be small. The largest imbalances occur when the  $L_t$  have the same sign, a circumstance in

<sup>3</sup> This probability is equal to the area of the set  $\{(L_1, L_2) \mid -L < L_1 < L, -L < L_2 < L, |aL_2 + a^3L_1| > |aL_2 + a^2L_1|\}$

divided by  $L^2$ .

which the retrospective iteration always reduces the imbalance. Thus, the retrospective iteration has the desirable characteristic that it tends to be useful when imbalances are potentially large. This effect can be measured quantitatively in our one-sector example. Assume that the  $L_t$  are independent random variables with mean zero and a common variance  $\sigma_L^2$ . If we assume in addition that the loss caused by an imbalance is proportional to the square of the magnitude of the imbalance, then it follows that the expected value of the loss is proportional to the variance of the imbalance.

We have

$$\begin{aligned}\sigma^2(E_2) &= \sigma^2[aL_2 + a^3 L_1] = a^2 \sigma_L^2 + a^6 \sigma_L^2 \\ &= (a^2 + a^6) \sigma_L^2\end{aligned}$$

and

$$\begin{aligned}\sigma^2(E'_2) &= \sigma^2[aL_2 + a^2 L_1] = a^2 \sigma_L^2 \\ &= (a^2 + a^4) \sigma_L^2\end{aligned}$$

Hence

$$\frac{\sigma^2(E_2)}{\sigma^2(E'_2)} = \frac{a^2 + a^6}{a^2 + a^4} = .85, \text{ when } a = .5$$

Thus, the expected value of the loss caused by the supply-demand imbalance given a centrally planned iteration and a retrospective iteration is 85 percent as large as the expected value of the loss with a centrally planned iteration alone.

Because many commodities are not centrally planned, we are understating the value of the retrospective iteration by considering only its marginal contribution to the balancing process with the use of central planning (an external iteration) assumed to be given. Let us redefine  $E_2$  to denote the supply-demand imbalance in period 2 generated by a procedure which uses the retrospective iteration ( $\Omega = I, \Gamma = 0$ ); and let  $E'_2$  denote the imbalance when no iterations are used ( $\Omega = 0, \Gamma = 0$ ). Then, we have

$$E_2 = L_1 + aL_2$$

and

$$E'_2 = L_1 + L_2$$

In this case, given our previous assumptions about the losses caused by imbalances, and about the  $L_t$  and  $a$ , we have  $\sigma^2(E_2)/\sigma^2(E'_2) = .625$  so that the expected value of the loss caused by the supply-demand imbalance given a retrospective iteration is 62.5 percent of the expected loss with no iterations. If we repeat our calculations for  $E_4$  and  $E'_4$ , the respective imbalances after 4 periods, it turns out the expected loss from imbalances generated with the retrospective iteration is only 33.3 percent of the expected loss from imbalances generated without it.

These results lead me to conclude that in spite of the genuine element of risk that Montias has demonstrated, imbalances are likely to prove much less costly when the retrospective iteration is used than when it isn't.

One final point: the Montias argument is worrisome only to the extent that we believe that the  $L_t$  will differ in sign from one another. But given the fact that Soviet-type economies usually operate under very taut conditions, it seems likely that the  $L_t$  would usually have the same sign—a negative one.

## II. Planning Iterations and Capacity Limitations

Montias contends that the use of the retrospective iteration is likely to generate output targets that exceed capacity limitations. His argument rests on the supposition that  $Q_t$ , the exogenously determined increment to gross output, will be set by the planners to equal the additions to capacity accruing in period  $t$ . He argues further that at least some of the actual output levels for period  $t-1$  will have equaled capacity limits of that period. It follows, in his reasoning, that in many sectors  $Q_t$  will be set equal to the difference between the projected capacity limits for period  $t$  and  $Y_{t-1}$ , the actual outputs of the previous period. I agree that this may well be the case if planners had no intention of using the retrospective iteration, since in that situation current output targets are derived by adjusting actual outputs of the previous year. But if planners do intend to use the retrospective iteration it makes much more sense to assume that  $Q_t$  (for both lo-

cally and centrally planned commodities) would be set equal to the difference between projected capacity limits for period  $t$  and  $X_{t-1}$ , the gross demand of the previous year. (The planners would, after all, be aware that with a retrospective iteration, output targets are based on the previous year's gross demand.) Under the new assumption, there is no reason to believe that the retrospective iteration is more likely to generate infeasible output targets than any other type of iteration.

Unfortunately, there are some situations in which any type of planning iteration might generate output targets that exceed capacity limitations. In particular, if the final demand  $D_t$  has been specified, and if the gross output that is consistent with  $D_t$ , namely,  $(I - A)^{-1}D_t$ , significantly exceeds capacity limitations, then it is not possible to have a plan which is both reasonably consistent and feasible. If enough iterations are used to achieve a substantial degree of consistency then at least one of the iterations will generate output targets which exceed capacity limitations.

Montias suggests that this could be avoided by setting targets equal to production capacities in capacity-limited sectors and allowing net output quantities in those sectors (and related sectors) to be determined as residuals. I do not doubt that the Soviet planners actually engage in this practice to some extent. Nevertheless, this method of planning suffers from two shortcomings that would seriously limit its usefulness in many sectors of an economy. First, there is no way of knowing in advance how desirable the vector of net outputs will be. Secondly, when the supply of a particular net output turns out to be too low, it may not be easy to determine which sectors' capacity limitations caused that net output to be constrained to its low level.

The following alternative approach to the capacity problem is in the spirit of the model. Final demand,  $D_t$ , and the exogenous increment to gross output,  $Q_t$ , are selected with capacity constraints in mind. The scheduled planning iterations are then carried out. In some sectors the output targets generated by

the planning process exceed capacity limitations. This signals the planning authorities that the capacities of those sectors ought to be increased, or that the use of alternate technologies which require less of the scarce resource ought to be encouraged. But such changes cannot be brought about at once. In the short run, planned output targets which exceed capacity limits would have to be revised and set equal to those limits. Resulting shortages could be relieved by drawing on reserves, by substituting less scarce commodities for scarcer ones, and, if need be, by rationing.<sup>4</sup>

The advantage of this method of coping with capacity limitation problems is that it uses the material supply planning procedure to expose bottlenecks in the economy, so that the planners can correct them before they become serious.<sup>5</sup> This approach was tested as part of a 17-sector simulation model of a Soviet-type economy.<sup>6</sup> In the simulations, at least, it worked rather well.

#### REFERENCES

- H. S. Levine, "The Centralized Planning of Supply in Soviet Industry," in Joint Economic Committee, U.S. Congress, *Comparisons of the United States and Soviet Economies*, Washington 1959, 151-76.
- M. Manove, "A Model of Administrative Planning and Plan Execution in Soviet-Type Economies," unpublished doctoral dissertation, M.I.T. 1970.
- , "A Model of Soviet-Type Economic Planning," *Amer. Econ. Rev.*, June 1971, 61, 390-406.
- , "Non-Price Rationing of Intermediate Goods in Centrally Planned Economies," *Econometrica*, forthcoming.
- V. G. Treml, "The 1959 Soviet Input-Output Table (as Reconstructed)," in Joint Economic Committee, U.S. Congress, *New Directions in the Soviet Economy*, Part II-A, Washington 1966, 259-70.

<sup>4</sup> See my forthcoming paper.

<sup>5</sup> Herbert Levine quotes a Soviet source on exactly this point, see p. 162.

<sup>6</sup> See ch. 4 of my doctoral dissertation.

# The Determinants of U.S. Direct Investment in the E.E.C.: Comment

By MURRAY A. GOLDBERG\*

A variety of theories have been proposed to explain both the pattern and magnitude of U.S. direct foreign investment. One approach has been to explain direct foreign investment in terms of host country characteristics such as Gross National Product (*GNP*), rate of growth of *GNP*, trade barriers, and convertibility of currency. In a recent issue of this *Review*, Anthony Scaperlanda and Lawrence Mauer (S-M) adopt this approach in an effort to specify the determinants of U.S. direct investment in the European Economic Community (*E.E.C.*) over the period 1952-66.

S-M find that only the size of the *E.E.C.* market is a significant variable in explaining U.S. direct investment flows. In contrast this paper contends that the investment/size-of-market relationship that S-M test is theoretically questionable and that investment can be largely explained not by the *size* of the *E.E.C.* market but by the *growth* of that market.

Section I of this paper specifies a model based on the hypotheses proposed by S-M. This model is also applied to U.S. direct investment in the *E.E.C.* and the empirical results are discussed. Section II evaluates the S-M model and their empirical tests and compares their findings to those presented here. Section III offers a critique of the general method of studying direct foreign investment in terms of host country characteristics and suggests an alternative approach.

## I. A Model of Direct Foreign Investment

S-M suggest hypotheses in each of the

three categories of growth, tariff discrimination, and size of the market.<sup>1</sup>

The growth hypotheses "are fundamentally based on the relation between the level of aggregate demand and the stock of capital (total investment) needed to satisfy this demand" (p. 560). As aggregate demand increases, there is a flow of direct foreign investment—represented by *I*—to support a higher level of output. S-M measure aggregate demand by the size of *E.E.C. GNP*—represented by *Y*. Accordingly, *I* measure growth by  $\Delta Y$ , the absolute annual change in *E.E.C. GNP*, and hypothesize a positive relationship between investment and growth. This relationship between investment and the change in *GNP* is "similar to that expressed by the incremental capital-output ratio" (p. 560).<sup>2</sup>

The tariff discrimination hypothesis "is, that to avoid obstacles to trade . . . foreign investment is undertaken in the countries to which it is difficult to export because of the obstacle" (p. 561). As trade barriers increase, exports become less attractive and foreign investment becomes more attractive. In

<sup>1</sup> The reader is referred to Scaperlanda and Mauer's 1969 article for a detailed discussion of each hypothesis and references to the literature. All citations are from the 1969 article unless otherwise noted.

<sup>2</sup> There is perhaps a problem of statistical identification because of the "multiplier" relationship between investment and *GNP*. The direct investment flows studied here are defined as the annual change in the book value of U.S. direct foreign investment. As such they are not identical to "investment" in the sense of fixed capital formation, though they may give rise to autonomous expenditures which affect *GNP*. This effect, however, should be relatively minor. In 1962, the aggregate flow of direct foreign investment was on the order of 1½ percent of the gross, nonresidential, fixed capital formation of the *E.E.C.*, while expenditures on plant and equipment by U.S. affiliates in the *E.E.C.* were on the order of 2½ percent of the gross, nonresidential, fixed capital formation. These figures are based on B. Mueller and the *Survey of Current Business*.

\* Instructor in finance at the University of Illinois, Chicago Circle, and doctoral candidate in the Graduate School of Business of the University of Chicago. Helpful comments from Professors Robert Aliber and John Gould are gratefully acknowledged.

order to measure relative trade barriers, I define two ratios  $M_w$  and  $M_d$  as:

$M_w =$

$$\frac{\text{World exports to the } E.E.C. \\ \text{less } U.S. \text{ and } E.E.C. \text{ exports to the } E.E.C.}{E.E.C. \text{ exports to the } E.E.C.}$$

and  $M_d =$

$$\frac{\text{Developed Country}^3 \text{ exports to the } E.E.C. \\ \text{less } U.S. \text{ and } E.E.C. \text{ exports to the } E.E.C.}{E.E.C. \text{ exports to the } E.E.C.}$$

where the denominator is the exports of *E.E.C.* countries to other *E.E.C.* countries. Each of these ratios—similar to the ratio  $M$  defined by S-M—is “taken to serve as a proxy for the influence of obstacles on *U.S.* direct investment” and “is based on the assumption that increased effective discrimination will decrease imports from suppliers outside the discriminating area, while simultaneously increasing intra-area imports” (p. 562).

The tariff discrimination hypothesis suggests that the flow of investment is negatively related to  $\Delta M_w$  and  $\Delta M_d$ , a decline in  $M_w$  and  $M_d$  implying greater trade discrimination. One of several alternative hypotheses, however, is that in at least some industries, exports and foreign investment should be viewed as complements, rather than as substitutes, because of the tendency for *U.S.* firms to channel exports through foreign affiliates both for resale abroad and further processing.<sup>4</sup> In such cases higher trade barriers may discourage rather than encourage

direct investment. There is, therefore, no a priori expectation as to the sign of the relationship between  $I$  and  $\Delta M_w$  and  $\Delta M_d$ .

The size-of-market hypothesis proposed by S-M is discussed fully in Section II and is rejected there as untestable because the theoretical relationship between investment and the size of the market is undefined. The hypothesis is, consequently, not included in the model developed here.

Rather, the model I test is of the form

$$I = a_0 + a_1 \Delta Y + a_2 \Delta M'$$

where  $\Delta M'$  stands for  $\Delta M_w$  and  $\Delta M_d$ . The expected sign of  $a_1$  is positive, while the expected sign of  $a_2$  is uncertain. The model is also tested without the tariff discrimination proxy.

I employ two variations of  $I$  in the empirical tests;  $I_a$  and  $I_m$  are defined, respectively, as the annual change in the book value of aggregate and manufacturing *U.S.* direct investment in the *E.E.C.* Scaperlanda-Mauer suggest that aggregate *U.S.* direct investment offers the advantages of reducing specification errors and canceling transitory elements.<sup>5</sup> However, in view of the wide diversification of *U.S.* investment in the *E.E.C.* with flows for a typical year breaking down roughly as 60 percent manufacturing, 20 percent petroleum, 10 percent trade, and 10 percent other, it is also useful to study the flow of direct foreign investment in manufacturing alone. This is especially so since incremental capital-output ratios and tariff barriers seem most appropriate to manufacturing.<sup>6</sup>

Table 1 presents the regression results for both  $I_a$  and  $I_m$ . The data employed and their sources and units are given in the Data Appendix.

The regression equations offer relatively

<sup>3</sup> Developed countries are defined by the United Nations for these purposes as the United States, Canada, Western Europe, Australia, New Zealand, South Africa, and Japan.

<sup>4</sup> Raymond Mikesell, for example, has pointed out that “[a]bout one-fourth of all *U.S.* exports in 1964 went to *U.S.* affiliates abroad and nearly half of these were sold without further manufacturing. In 1964, 32 percent of *U.S.* exports of manufactures to Europe (excluding aircraft, mineral fuels, and processed agricultural products) went to *U.S.* affiliates” (pp. 445–46).

<sup>5</sup> The S-M argument as to the reduction of specification errors is based on a Yehuda Grunfeld-Zvi Griliches article cited by them. The Grunfeld-Griliches finding has, however, been seriously challenged by the more recent work of J. B. Edwards and Guy Orcutt.

<sup>6</sup> Since the data for *GNP* and exports cover all industries, the results derived for manufacturing investment flows can be considered only approximate.

TABLE 1—REGRESSION EQUATIONS RELATING AGGREGATE AND MANUFACTURING U.S. DIRECT FOREIGN INVESTMENT FLOWS TO THE *E.E.C.* TO SELECTED VARIABLES, 1952-1966

Dependent Variable	Independent Variables				$\bar{R}^2$	$S_e$	DW
	Constant	$\Delta Y$	$\Delta M_w$	$\Delta M_d$			
(1.1) $I_a$	-370.488 (3.258)**	52.889 (7.794)**			.810**	158.942	1.194**
(1.2) $I_a$	-241.784 (1.770)	51.633 (7.939)**	1481.76 (1.542)		.828**	151.138	1.363**
(1.3) $I_a$	-393.605 (2.925)*	52.919 (7.531)**		-772.179 (.335)	.797**	164.570	1.191
(1.4) $I_m$	-233.496 (2.870)*	32.104 (6.614)**			.753**	113.697	1.788*
(1.5) $I_m$	-153.828 (1.536)	31.326 (6.572)**	917.208 (1.302)		.766**	110.772	2.063*
(1.6) $I_m$	-311.801 (3.680)**	32.206 (7.278)**		-2615.68 (1.909)	.794**	103.639	1.744*

Sources and Units: See Data Appendix.

Notes: The numbers in parentheses are *t*-statistics. For the regression coefficients and the  $\bar{R}^2$ , \* and \*\* indicate significantly different from zero at the 5 percent and 1 percent levels, respectively.

For the Durbin-Watson statistic, \* and \*\* indicate that the null hypothesis of residual independence cannot be rejected at the 5 percent or 1 percent level of significance, respectively. This test is based on the Thiel-Nagar procedure.

good fits for both aggregate and manufacturing investment flows. The independent variable  $\Delta Y$  is highly significant and properly signed in all equations. The tariff discrimination proxy, on the other hand, offers mixed results. The coefficient of  $\Delta M_w$  is positively signed in both cases; and though not significant at the 5 percent level, it does contribute explanatory power as reflected in the increase in adjusted  $\bar{R}^2$  when it is present. The coefficient of  $\Delta M_d$  is negatively signed in both cases. It contributes nothing to an understanding of aggregate investment flows and there is evidence of serial correlation in the aggregate investment equation. In the case of manufacturing investment flows, however,  $\Delta M_d$  is significant at the 10 percent level and the highest  $\bar{R}^2$  is obtained when it is included.

The growth hypothesis is thus supported by the model developed here. The regression results imply that about 80 percent of aggregate investment flows and about 75 percent of manufacturing investment flows may be viewed as an adjustment in the stock of

existing investment so as to satisfy a higher level of demand, as reflected in *GNP*. The findings are, on the other hand, inconclusive with respect to both the sign and the significance of the relationship between investment flows and trade barriers.

## II. A Review of the S-M Findings

Scaperlanda-Mauer test all three categories of their hypotheses through a variety of variables. To test the growth hypothesis, they include beside  $\Delta Y$ , both  $G_1$ , the percentage rate of growth of *E.E.C. GNP*, and  $G_2$ , the ratio of  $G_1$  to the percentage rate of growth of *U.S. GNP*. As S-M seem to recognize, however, such a relationship between the *dollar* flow of investment and the *percentage* rate of growth or relative growth is open to question. In both 1955 and 1963, for example, *E.E.C. GNP* grew by about 10½ percent. In the former case this represented an increase of \$11.68 billion, while in the latter case this represented an increase of \$24.07 billion. In 1955 the flow of aggregate invest-

ment was \$152 million; in 1963 it was \$768 million. The same *percentage* rate of growth thus led to quite different *dollar* flows of investment in the two years.

S-M test the tariff discrimination hypothesis by defining a proxy  $M$  as:

$$M = \frac{\text{U.S. exports to the E.E.C.}}{\text{E.E.C. exports to the E.E.C.}}$$

This proxy  $M$ , however, is marred on both theoretical and data grounds. If exports and subsidiary production are substitutes for one another (as S-M imply), then investment and exports will vary together independent of trade discrimination. For example, assume an American firm  $XYZ$  exporting some product to the *E.E.C.* By 1960 its exports have grown to \$1 million annually and it decides that the level of sales justifies establishing a production subsidiary. This is effected by an investment of \$10 million and the firm ceases to export to the *E.E.C.* In 1960, then, the flow of investment to the *E.E.C.* includes  $XYZ$ 's \$10 million, while *U.S.* exports to the *E.E.C.* decline by \$1 million. The proxy  $M$  declines not because there has been a change in trade discrimination but because there has been an autonomous flow of investment. Thus, even if  $I$  and  $\Delta M$  are inversely related because investment and exports are substitutes, the direction of causation is undetermined.

The data on which  $M$  is based also present a problem because the values for *U.S.* exports to the *E.E.C.* are internally inconsistent. The difficulty is that the "special category" exports of the United States are included in the trade figures for 1955 and beyond but are excluded from the earlier values.<sup>7</sup>

It is to overcome both these difficulties that I employ  $M_w$  and  $M_d$  in this study. If the obstacles to trade with the *E.E.C.* facing

all countries or all developed countries are similar and if the commodity composition of their exports to the *E.E.C.* is similar, then  $M_w$  and  $M_d$  should prove to be better proxies for trade barriers than  $M$ .<sup>8</sup>

S-M also test the tariff discrimination proxy in both the  $M$  and  $\Delta M$  form. Yet there is no reason why any given level of  $M$  should give rise to a flow of investment. Rather, investment should flow in response to a *change* in the set of obstacles to trade.

For S-M "... the size-of-market hypothesis is that foreign investment will take place as soon as the market is large enough to permit the capturing of economies of scale" (p. 560). The size of the market is measured by  $Y$ , *E.E.C. GNP*, and it is assumed that as *GNP* increases new opportunities for scale economies emerge and foreign investment is attracted. S-M seek to relate  $I$  directly to  $Y$  and hypothesize that the higher the *level* of *GNP* the greater will be the *flow* of direct investment. But the fact that investment becomes profitable when *GNP* achieves some critical value which allows economies of scale to be realized does not in any way indicate what the *magnitude* of such investment will be. A positive association between the *flow* of new investment and the *level* of *GNP* presumes either that larger and larger *GNP* allows economies of scale to be enjoyed by an ever-growing number of firms, so that the aggregate flow of new investment continues to rise, or that larger and larger *GNP* implies larger initial investments. There is little basis for either presumption.<sup>9</sup> Theoretically, the relationship between the flow of investment and the level of *GNP* on the grounds

<sup>8</sup> Actually, non-*U.S.* trade with the *E.E.C.* will also be affected both by non-*U.S.* investment in the *E.E.C.* and *U.S.* investment in the *E.E.C.*, but these effects should be small in comparison with the effects of *U.S.* investment on *U.S.* exports.

<sup>9</sup> For example, firms  $ABC$  and  $XYZ$  may both be waiting for *GNP* to achieve a size such that they can profitably invest \$10 million and begin to produce their products abroad. That critical size may be \$100 billion for firm  $ABC$  and \$200 billion for firm  $XYZ$ . In each case the flow associated with the two levels of *GNP* is identical. Investment flows will increase with *GNP* only if, *ceteris paribus*, an increasing number of firms make investments or increasingly larger investments are made.

<sup>7</sup> As described by P. J. Loftus, Director of the United Nations Statistical Office, this arises because "... prior to 1965, the United States data referred to national exports and the special category exports of the United States were not available by destination. In 1965 we were able to show from 1955 onwards, the special trade data, as well as to distribute the special category exports of the United States by regions of destination."

of economies of scale could be of any form whatsoever.

Furthermore, the economies-of-scale hypothesis implies that if *GNP* did not increase one year there would be no new investment on grounds of economies of scale, since no new opportunities for scale economies would have arisen. Thus if  $Y_t = Y_{t-1}$  (where  $t$  subscripts refer to time), presumably  $I_t = 0$ . Yet the nature of the relation tested by S-M is such that if  $Y_t = Y_{t-1}$ ,  $I_t = \alpha + \beta Y_t$  (ignoring other influences). The contradiction arises because the underlying theory says only that some new investments become profitable at higher levels of *GNP* but gives no indication of the size of such investments. The S-M equation, however, relates the magnitude of the flow of investment to the level of *GNP*.

For these reasons I make no attempt to incorporate the size-of-market hypothesis into the model developed in Section I.

Finally, the S-M data for  $I_a$  (their  $I$ ) and *E.E.C.* exports to the *E.E.C.* contain a number of errors, as may be seen by comparing their Appendix of Basic Data to the Data Appendix given here.<sup>10</sup>

Despite the variety of S-M's variables, their study yields only one significant relationship, that between  $I$  and  $Y$  with an  $\bar{R}^2$  of .959 (1971, p. 509).<sup>11</sup> That relationship offers a better fit than any of the equations tested here but is statistically suspect. By focusing on *levels* of investment and *GNP* rather than on *changes* in those variables, S-M introduce into their regression the problem of the tendency for both investment and *GNP* to increase over time. The importance of this common time trend in contributing to the high  $\bar{R}^2$  can be demonstrated by relating the first differences of  $I$  and  $Y$ . A regression of  $\Delta I$  on  $\Delta Y$  yields the much less impressive

$\bar{R}^2$ 's of .270 for the S-M data and .212 for the corrected data.<sup>12</sup> The S-M results are thus largely only a reflection of the common time trend in both investment and *GNP*. S-M nonetheless conclude "that only the size-of-market hypothesis can be supported statistically. Negative findings were discovered for all variants of growth and tariff discrimination hypotheses; these hypotheses were rejected as not statistically significant regardless of the model and time period studied" (pp. 566-67).<sup>13</sup>

In contrast, I have demonstrated that—when the size-of-market hypothesis is deleted on both theoretical and statistical grounds—the growth hypothesis yields highly significant results in at least one of its formulations for both aggregate and manufacturing investment, while the tariff discrimination hypothesis offers inconclusive results.<sup>14</sup>

*GNP* is the only significant variable in the S-M analysis; yet it enters into consideration solely on the basis of the size-of-market hypothesis based on economies of scale—a somewhat shaky framework at best. This perhaps says more about the analysis than about the variable.

### III. A Critique of the Approach

One of the fundamental shortcomings of a model of U.S. direct foreign investment based on host country characteristics is that although host country characteristics may explain why the host region is an attractive area for investment, they cannot explain why the investment originates in the United States, rather than in the host region itself or some other country.

The variables suggested by S-M such as growth, tariff discrimination, and the size of

<sup>10</sup> Some of the errors arise from S-M's use of preliminary figures.

<sup>11</sup> When the S-M equations are reestimated using the corrected data, the results are fundamentally unchanged and *GNP* ( $Y$ ) remains the only significant variable. In all cases, however, the  $\bar{R}^2$ 's are lower, while the standard errors and Durbin-Watson statistics are larger. In particular, for the equation relating  $I$  to  $Y$ , the  $\bar{R}^2$  drops to .955 and the standard error rises from 74.022 to 77.550. The coefficient for  $Y$  also declines from 4.900 to 4.885.

<sup>12</sup> Both  $\bar{R}^2$ 's are significant at the 5 percent level. For the equation using S-M data the constant has a value of  $-72.217$  and a  $t$ -statistic of 1.107 and  $\Delta Y$  has a coefficient of 9.628 and a  $t$ -statistic of 2.474. The values for the equation using corrected data are similar.

<sup>13</sup> S-M test various combinations of the size-of-market, growth, and tariff discrimination variables. They also run regressions covering three time spans: 1952-66, 1952-58, and 1959-66.

<sup>14</sup> The use of  $\Delta Y$  in the model tested here eliminates much of the problem associated with the common time trend in  $I$  and  $Y$ .

the market are basically independent of any attribute of the United States. The model provides no insight into the nature of the comparative advantage of U.S. firms that allows them to compete effectively in foreign markets.

There are, also, other discomfiting aspects to this approach to the study of direct foreign investment. Even though empirical work—such as that of S-M and that reported here—is often conducted at a highly aggregated level, it rarely incorporates those factors and those models which are useful in understanding aggregate *domestic* investment. These include expected profits, liquidity, cost of capital, and the flexible accelerator.<sup>15</sup> All these, plus the interrelationship between U.S. domestic and foreign investment, are often neglected, as is the concept of investment in a stock-adjustment framework.<sup>16</sup> For authors like S-M direct investment is an instantaneous adjustment to changed conditions.

At its basis the phenomenon of direct foreign investment is distinguished from that of domestic investment by two principle considerations. First, the foreign firm incurs costs that a host country firm does not. These are in the form of political and exchange risk and in the form of a penalty for being foreign and thereby having less familiarity with the market and generally longer lines of communication and control. To compensate for these additional costs, the foreign firm must possess some "advantage" over host country firms that allows it to nonetheless compete effectively.<sup>17</sup> Second, the foreign firm has the choice of serving the market through exports or through the production

of a subsidiary.<sup>18</sup> The decision to invest abroad implies that the cost of producing abroad is below that of the landed cost of exporting when all costs, such as those of production, sales, and administration, are included.

A study which seeks the "determinants of U.S. direct foreign investment" must encompass both these considerations by specifying the nature of the advantage that allows the foreigner to compete abroad, the degree of penetration of the market that the advantage permits, and the portion of foreign sales that will be serviced by foreign production.

#### REFERENCES

- D. Devlin and F. Cutler, "The International Investment Position of the United States: Developments in 1968," *Surv. Curr. Bus.*, Oct. 1969, 49, 23-36.
- J. B. Edwards and G. H. Orcutt, "Should Aggregation Prior to Estimation Be the Rule," *Rev. Econ. Statist.*, Nov. 1969, 51, 409-20.
- D. W. Jorgenson and C. D. Siebert, "A Comparison of Alternative Theories of Corporate Investment Behavior," *Amer. Econ. Rev.*, Sept. 1968, 58, 681-712.
- C. P. Kindleberger, *American Business Abroad*, New Haven 1969.
- W. Lederer and F. Cutler, "International Investments of the United States in 1966," *Surv. Curr. Bus.*, Sept. 1967, 47, 39-51.
- R. F. Mikesell, "Decisive Factors in the Flow of American Direct Investment to Europe," *Econ. Int.*, Aug. 1967, 20, 413-53.
- B. Mueller, *A Statistical Handbook of the North Atlantic Area*, New York 1965.
- E. Nelson and F. Cutler, "The International Investment Position of the United States in 1967," *Surv. Curr. Bus.*, Oct. 1968, 48, 19-32.
- S. Pizer and F. Cutler, "Foreign Investments in 1963-64," *Surv. Curr. Bus.*, Aug. 1964, 44, 8-14, 24; "1964-65," Sept. 1965, 45, 22-32; "1965-66," Sept. 1966, 46, 30-40.
- and ——— "U.S. International Invest-

<sup>15</sup> See, for example, Dale Jorgenson and Calvin Siebert.

<sup>16</sup> In contrast, for example, Alan Severn has made preliminary findings to the effect that the methods of investigation typically applied to domestic investment are also applicable to foreign investment and that foreign and domestic investment are interrelated through the financing mechanism that allocates internally generated funds among investment alternatives.

<sup>17</sup> Examples of such an advantage suggested in the literature are patented technology, a differentiated product, special marketing skills, and a monopolistic position arising from imperfections in markets for goods and factors. See, for example, C. P. Kindleberger.

<sup>18</sup> The domestic firm also has the choice of investing and producing its own goods or importing them. For the United States especially, however, this is a minor consideration.

- ments," *Surv. Curr. Bus.*, Aug. 1963, 43, 16-22, 28.
- and ——— "Expansion in U.S. Investments Abroad," *Surv. Curr. Bus.*, Aug. 1962, 42, 18-24, 32.
- and ——— "United States Assets and Investments Abroad," *Surv. Curr. Bus.*, Aug. 1961, 41, 20-26.
- and ——— "United States Foreign Investments: Measures of Growth and Economic Effects," *Surv. Curr. Bus.*, Sept. 1960, 40, 15-24.
- and ——— "Capital Flow to Foreign Countries Slackens," *Surv. Curr. Bus.*, Aug. 1959, 39, 25-32.
- and ——— "Private Foreign Investments Near \$37 Billion," *Surv. Curr. Bus.*, Sept. 1958, 38, 15-23.
- and ——— "Record Growth of Foreign Investments," *Surv. Curr. Bus.*, Aug. 1957, 37, 22-30.
- and ——— "Growth of Foreign Investments in the United States and Abroad," *Surv. Curr. Bus.*, Aug. 1956, 36, 14-24.
- and ——— "International Investments and Earnings," *Surv. Curr. Bus.*, Aug. 1955, 35, 10-20.
- and ——— "Foreign Investments and Income," *Surv. Curr. Bus.*, Nov. 1954, 34, 6-13, 22-23.
- and ——— "Growth in Private Foreign Investments," *Surv. Curr. Bus.*, Jan. 1954, 34, 5-10.
- A. E. Scaperlanda and L. J. Mauer, "The Determinants of U.S. Direct Investment in the E.E.C.," *Amer. Econ. Rev.*, Sept. 1969, 59, 558-68.
- and ——— "Errata," *Amer. Econ. Rev.*, June 1971, 61, 509-10.
- A. K. Severn, "International and Financial Behavior of American Direct Investors in Manufacturing," paper presented to the Conference on International Mobility and Movement of Capital, sponsored by Universities-Nat. Bur. Com. Econ. Res., Jan. 30-Feb. 1, 1970.
- H. Thiel and A. L. Nagar, "Testing the Independence of Regression Disturbances," *J. Amer. Statist. Ass.*, Dec. 1961, 56, 793-806.
- International Monetary Fund, *International Financial Statistics*.
- United Nations, *Yearbook of International Trade Statistics*, New York 1960-62, 1964, 1966-69.
- , private correspondence from P. J. Loftus, Director of the Statistical Office, July 17, 1970.
- U.S. Department of Commerce, private correspondence from Seiko N. Wakabayashi, Acting Assistant Chief for Statistics, Balance of Payments Division, Office of Business Economics, July 21, 1970.

Data Appendix follows

DATA APPENDIX  
(dollars)

Year	$I_a$ (millions)	$I_m$ (millions)	$Y$ (billions)	U.S. Exports to the <i>E.E.C.</i> (millions)	<i>E.E.C.</i> Exports to the <i>E.E.C.</i> (millions)	World Exports to the <i>E.E.C.</i> (millions)	Developed Country Exports to the <i>E.E.C.</i> (millions)
1951			87.18	2050	4250	15090	10140
1952	68	-113	97.58	1770	4290	14400	9990
1953	98	188	103.87	1460	4590	14560	10000
1954	101	50	110.99	1810	5210	16480	11180
1955	152	62	122.67	2590	6210	19240	13550
1956	238	95	134.44	3650	7070	22490	16180
1957	281	172	146.66	3860	7880	24050	17740
1958	228	139	159.10	2840	7530	22150	15830
1959	300	180	171.72	2830	8400	23730	17060
1960	436	287	189.85	3930	10250	28150	20850
1961	460	263	206.45	4100	11900	30790	23310
1962	618	404	226.22	4530	13560	34110	26110
1963	768	424	250.29	4860	15920	38530	29650
1964	936	611	275.82	5230	18390	43080	33220
1965	878	586	299.42	5200	20820	46670	36130
1966	1280	676	321.62	5460	23230	51087	39460

*Sources.* Aggregate U.S. direct foreign investment flows ( $I_a$ ) and manufacturing U.S. direct foreign investment flows ( $I_m$ ) are based on the U.S. Department of Commerce's annual article in the *Survey of Current Business* on the value of U.S. direct foreign investment. The data for  $I_m$  have been supplemented by information obtained from Seiko N. Wakabayashi of the Office of Business Economics.

*E.E.C. GNP (Y)* is from Scaperlanda and Mauer who derive it from B. Mueller and *International Financial Statistics*.

U.S. exports to the *E.E.C.*, *E.E.C.* exports to the *E.E.C.*, World exports to the *E.E.C.*, and Developed Country exports to the *E.E.C.* are—with the exception of the 1951 values for the last three series—all from the United Nations based on the *Yearbook of International Trade Statistics* and private correspondence from P. J. Loftus, Director of the Statistical Office. The United Nations has revised the series for *E.E.C.* exports to the *E.E.C.* from 1952 onwards. The published value for 1951 is, therefore, not consistent with the remainder of the series presented here. I have, accordingly, revised the published value for 1951 by an amount equal to the United Nations revision for 1952. This, in turn, gives rise to a revision of equal amount in the 1951 values for World and Developed Country exports to the *E.E.C.*

# The Determinants of U.S. Direct Investment in the E.E.C.: Reply

By ANTHONY E. SCAPERLANDA AND LAURENCE J. MAUER\*

In order to reply to Murray Goldberg's comment, we shall identify its theoretical and empirical components. Goldberg's "theoretical" component, which is considered first, is essentially a series of *ad hoc* assertions, which provide little additional contribution to the emerging theory of direct foreign investment. The purpose of our article was to identify a number of hypotheses which have been either proposed or considered by the major contributors to the literature. From these existing categories of hypotheses we identified three categories which were deemed statistically testable. This process involved empirically specifying variables which were designed to capture the effects of selected relationships which had been postulated by others. With our major objective clearly the testing of existing hypotheses, we made no attempts to develop new hypotheses. At most, one should expect that empirical findings such as ours would lead to the refinement of previously existing theoretical propositions.

In his comment, Goldberg seems to be insufficiently aware of the theoretical underpinning of the proxies used in our empirical testing. An important case in point is the variable which we labeled the "tariff discrimination proxy" (although we mentioned in passing many obstacles to trade in addition to tariffs). Although this hypothesis has been widely dealt with in the literature (see: Bela Balassa (1961), p. 182; Richard Caves; CED, p. 35; Anthony Edwards, pp. 5-7; Randall Hinshaw, p. 103; F. B. Jensen and Ingo Walter, p. 237; Maynard Kohler, pp. 93-94; Lawrence Krause, p. 120; Robert

Mundell; Samuel Pizer and F. Cutler; Walter Salant, pp. 122-23; Andrew Schmitz and Peter Helmberger; D. A. Snider; Erich Spittaller; and Paul Streeten, p. 41),<sup>1</sup> he seems unaware that the overwhelming consensus in this literature does not support the specification which he employs. He arbitrarily replaced the numerator of our  $M$  with either world exports to the E.E.C. or developed countries exports to the E.E.C. (in both cases minus U.S. and E.E.C. exports to the E.E.C.).<sup>2</sup> Such redefinitions would be acceptable if one were concerned with either

<sup>1</sup> The fact that an author is cited is not meant to imply that he has accepted the "tariff discrimination hypothesis." Rather it is an indication that he has at least recognized the existence of the hypothesis in the form presented here. Also, some of the citations mentioned have appeared since the publication of our article (the same can be said for citations associated with the size of market hypothesis).

<sup>2</sup> At one point in his comment Goldberg justifies his alternative definitions of  $M$  on the grounds that our data are "internally inconsistent." Although his terminology would seem to over-state the problem, we have verified on rechecking that he is correct in his position that the data as reported understate the U.S.-E.E.C. trade flows for the first three years of our time-series. Although we did not anticipate that our conclusions would be affected by this data deficiency, we did retest the performance of our  $M$  using data free of the 1952-54 problems; this investigation gave no statistical support to the tariff discrimination hypothesis.

At another point he indicates that our E.E.C.-E.E.C. trade data and investment data "contain a number of errors." Generously he indicates in footnote 10 that: "Some of the errors arise from S-M's use of preliminary figures." Upon rechecking the data, it was found that every one of the few existing discrepancies between our data and Goldberg's resulted from the use of preliminary data. The resulting small discrepancies do not seem to have had an influence on the statistical tests.

While digressing on the subject of data, the reader should take note that the data which Goldberg presents in his Data Appendix for  $I_m$  for 1952-57 are incorrect. The corrected data are as follows: 1952—\$44 million; 1953—45; 1954—60; 1955—74; 1956—118; 1957—113. According to the U.S. Department of Commerce, the definitive data source for the pre-1961 is *Balance of Payments Statistical Supplement: Revised Edition*.

\* Associate professors of economics, Northern Illinois University. Preparation for this paper was partially financed by the National Science Foundation, Grant No. G.S. 2235. Thanks are also due to R. George for research assistance and to Julius N. Friedlin of the U.S. Department of Commerce for his helpful comments.

rest of the non-*U.S.* and non-*E.E.C.* world or non-*U.S.* and non-*E.E.C.* developed countries' investment in the *E.E.C.*<sup>3</sup> But such is not the concern of either our article or his comment. Given the emphasis on *U.S.* direct investment, an application of the hypothesis, as generally stated in the literature, requires that the substitutability of *U.S.* direct investment for *U.S.* exports (as set in perspective by including internal *E.E.C.* trade in the variable) be tested. Thus, accepting our basic definition of *M* (as Goldberg does), the hypothesis as formulated in the literature supports our original specification of *M* as superior to Goldberg's.

Finally, and probably most important, in his Section II, Goldberg claims to offer both theoretical and empirical grounds for excluding from the direct investment explanatory model the market size variable. Below it is shown that both his claims are incorrect. Indeed, it would seem difficult to deny the theoretical validity of the market size hypothesis which has been around at least since 1776, and which has in recent years been postulated in connection with the effects of a customs union's creation on direct foreign investment. (See Robert Aliber; Vladimir Bandera and J. T. White; Balassa (1966), (1967, p. 128); Caves; John Dunning, p. 299; Corwin Edwards; Lawrence Krause, p. 120; Richardson; Spittaller; and Guy Stevens).<sup>4</sup>

<sup>3</sup> Such redefinitions might also be acceptable if one could assume an identical product mix for *U.S.* exports to the *E.E.C.* and non-*U.S.*, non-*E.E.C.* exports from either the world or developed countries to the *E.E.C.* But such an assumption does not seem warranted especially in light of the fact that product differentiation permits substantial variation in the "product-mix" even within a given commodity class. Alternatively he might have some grounds for redefining the *M* proxy if *U.S.* direct investment were but a small fraction of total foreign investment which might be "encouraged" by Common Market oriented trade barriers. However, considering the dominant position of *U.S.* direct investment relative to total foreign investment in the *E.E.C.* this justification does not seem to exist.

<sup>4</sup> A survey of the theoretical implications of the size of market hypothesis could be obtained from several of these sources. For example, one might begin with Dunning, p. 299, who provides a general description of the hypothesis. Stevens, p. 139, among others, discusses the cost of initiating direct foreign investment. These costs can be justified only after the market has

While Goldberg claims to provide theoretic grounds for rejecting the market size hypothesis, his argument is not, in fact, directly concerned with the theoretical validity of the relationship between direct investment and market size. Instead the first line of reasoning which he sets forth is concerned with the validity of the particular functional form of this relationship, a question which we feel can only be resolved on empirical grounds. Given our judgment as to the nature of this problem, the question to which we addressed ourselves was specific in temporal and geographic terms: which of a set of suggested hypotheses explaining *U.S.* direct investment in the *E.E.C.* is empirically supported for the time period 1952-66? With this emphasis, the direct investment-market size relationship was initially tested in a linear nonhomogeneous form (as reported in our article); in addition, supplementary regressions (which were not reported) were run to examine alternative nonlinear functional forms for this relationship. The linear form appeared to yield the best fit in this empirical context.

The second line which Goldberg follows in attempting to dispose of the market size hypothesis is to assert on intuitive grounds that initial investment (that is investment in new ventures) is less important than expansion investment (which may, for example, be associated with the acceleration or absolute growth hypothesis which he advocates). Again, the empirical context for this evaluation is crucial. The study by Vaupel and Curhan (p. 10) based on a sample of 187 major *U.S.* manufacturing corporations, showed that during the period of our study the number of firms from the sample operating in the *E.E.C.* virtually doubled. In addition, there are several reasons to believe that expansion investment in some time periods and/or in some geographic areas may also

---

attained a critical minimum size sufficient to permit an optimum scale initial investment. Richardson, p. 92, has explored reasons why the accelerator influence (the *I-ΔY* relationship) may not be strong. Finally a geometric presentation of the size of market hypothesis has been provided by Aliber, p. 25.

be chiefly determined by market size in the direct investment case rather than by change in market size or an alternate growth concept. For example, market size considerations may be viewed as very important for investment associated with the production and marketing of new product lines, even though firms undertaking this investment might already have operations in the *E.E.C.* Interestingly, this relationship may be expected to be prevalent in direct investment following the reasoning of the monopolistic "advantage" framework suggested by Goldberg. The argument goes that when a firm has a monopolistic advantage (which may be temporary) resulting from access to information or knowledge linked to research and development activities, the direct investment of that firm may be undertaken explicitly to gain from this advantage. Thus new foreign investment (even if undertaken by an established firm) in the producing and marketing of new products can be expected, other things equal, in foreign countries as the market size in these countries becomes sufficient.

Another example in which market size considerations may be viewed as the chief determinants of expansion investment is the case of vertical direct investment. For this case, investment is undertaken as the size of market becomes sufficient to permit the inclusion of additional input processes in the operations located abroad. A number of investigators have noted that the vertical investment pattern is a common one in firms operating abroad. A typical example would have a sales subsidiary developed first to establish and test the market, followed (if the market is found to be of sufficient size) by subsequent backward integration. (See Behrman, pp. 63 ff; Caves, p. 3; Richardson, p. 91-92.) This integration would usually first involve investments that could be operated at optimal capacity given the estimated market size. If additional investments are desired, one would expect that they normally would not be undertaken until the market reached the minimal critical size to permit optimal utilization of the desired investment.

Beyond his theoretical assertions, Gold-

berg claims statistical grounds on which to reject the market size hypothesis. His argument is that investment and *GNP* are both dominated by a common time trend, and it is on this ground that he claims we find a high correlation between  $I$  and  $Y$ . If one grants this claim, he must also accept the possibility of trend influencing the statistical relationship between  $I$  and  $\Delta Y$ . For example, consider a situation in which a country is experiencing a constant percent growth per annum in *GNP*. Under these circumstances, the size of market proxy  $Y$  for this country would be trend dominated, but the associated  $\Delta Y$  variable too would be trend dominated. In order to test whether the trend factor makes a difference in the conclusions which we set forward in our original study and in Goldberg's comment, we have adopted the procedure of detrending variables by regressing them on time.<sup>5</sup>

The residuals from each relationship were then regressed following the models presented in our earlier article and in Goldberg's comment, with no change in our earlier conclusions for all regression models run in our earlier study. For brevity and to confront Goldberg's position, the results with regard to the market size ( $Y$ ) and absolute growth ( $\Delta Y$ ) hypotheses are presented in Table 1, where lower case letters are used to identify the residuals from the appropriate variable-time relationships corresponding to the capital letters in our original variable designation.

Using all detrended variables, support is shown for the market size hypothesis, while no support is given the absolute growth hypothesis, even in the regression model (1.2) which Goldberg prefers. For equations (1.1) and (1.3), rather high  $\bar{R}^2$ 's are seen, and there is no evidence of serial correlation; for these equations, the variable  $y$  is significant at the 1 percent level.

In summary, we do not believe that Goldberg is justified in eliminating from his analysis the size of market hypothesis; to do

<sup>5</sup> For an example of this procedure see Charles Schotta. The  $\bar{R}^2$ 's for these regressions on time were .87, .97, and .88 for  $I$ ,  $Y$ , and  $\Delta Y$ , respectively.

TABLE 1—REGRESSION EQUATIONS USING RESIDUALS FOR U.S. DIRECT FOREIGN INVESTMENT IN THE E.E.C.: ANNUAL DATA, 1952-1966<sup>a</sup>

Regression Equation	$\bar{R}^2$	$Se$	$DW$
(1.1) $i = .000 + 8.10y$ (6.27)**	.752**	63.46	2.50*
(1.2) $i = .000 + 12.5 \Delta y$ (.84)	.051	124.04	.87
(1.3) $i = .000 + 8.12y - .444\Delta y$ (5.82)** (.05)	.752**	66.04	2.60**

<sup>a</sup> The numbers in parentheses are *t*-statistics; \* and \*\* indicate significantly different from zero at 5 percent and 1 percent levels, respectively.

The symbol \* or \*\* on the Durbin-Watson statistics indicates that the null hypothesis of residual independence cannot be rejected at the 5 percent or 1 percent level of significance. The tests of the *DW* statistics are based on the H. Theil-A. Nager testing procedure.

so is both theoretically unsound and an empirical misspecification of the direct investment model.

Turning to the empirical work presented in Goldberg's comment, we see that by focusing on direct investment in manufacturing rather than on aggregate direct investment, he introduces a new and potentially significant dimension to the empirical literature. Unfortunately many specification problems beset his analysis. First, the empirical model used to test the determinants of manufacturing direct investment omits a proxy for the market size variable. In light of our earlier findings and our analysis above, it is likely that serious errors in specification arise from omitting this variable. Second, as argued above, the Goldberg specification of the tariff discrimination variable is incorrect; therefore his test of the tariff discrimination hypothesis should not be taken seriously.<sup>6</sup> The last problem which enters Goldberg's empirical work arises from his not developing the manufacturing dimension for both dependent and independent variables. While his dependent variable relates to man-

ufacturing alone, his independent variables are developed from *GNP* and exports for all industries. For example, in the manufacturing equations, *M* should be specified to include only trade in manufactured products.

In summary, our central point is that for both theoretic and empirical reasons the direct investment model should incorporate a proxy for the influence of size of market on direct investment. A second point is that proxies for this and other influences must be defined so that both the variable specification and the data used properly reflect the theoretical underpinnings of the hypotheses being tested. Third, implicit (and at points explicit) in both our original article and this reply is the fact that both geographic and temporal dimensions of the analysis exist and must be recognized in the process of empirical hypotheses testing. It is conceivable that with regard to a different geographic area, the determinants of direct foreign investment may be somewhat different from those identified in the *E.E.C.* It is also conceivable that additional determinants may come to be supported statistically as time passes, though this judgment fundamentally must be made on empirical grounds.

#### REFERENCES

- R. Z. Aliber, "A Theory of Direct Foreign Investment," in C. P. Kindleberger, ed., *The*

<sup>6</sup> Even within his own incomplete model, if he had correctly specified *M*, he would have found less support for the absolute growth hypothesis and greater support for the tariff discrimination hypothesis. Our position is based on estimates using 1955-69 data rather than the original 1952-66 series.

- International Corporation*, Cambridge, Mass. 1971.
- V. N. Bandera and J. T. White, "U.S. Direct Investments and Domestic Markets in Europe," *Econ. Int.*, Feb. 1968, 21, 117-33.
- B. Balassa, "American Direct Investments in the Common Market," *Banca Naz. Lavoro Quart. Rev.*, June 1966, 121-46.
- , *The Theory of Economic Integration*. Homewood 1961.
- , *Trade Liberalization Among Industrial Countries*, New York 1967.
- J. N. Behrman, "Foreign Associates and Their Financing," in R. F. Mikesell, ed., *U.S. Private and Government Investment Abroad*, Eugene 1962.
- R. E. Caves, "International Corporations: The Industrial Economics of Foreign Investment," *Economica*, Feb. 1971, 38, 1-27.
- J. H. Dunning, *Studies in International Investment*, London 1970.
- A. Edwards, *Investment in the European Economic Community*, New York 1964.
- C. D. Edwards, "Size of Markets, Scale of Firms, and the Character of Competition," in E. A. G. Robinson, ed., *Economic Consequences of the Size of Nations*, New York 1960.
- R. Hinshaw, *The European Community and American Trade*, New York 1964.
- F. B. Jensen and I. Walter, *The Common Market: Economic Integration in Europe*, New York 1965.
- M. Kohler, *The Common Market and Investment*, New York 1960.
- L. B. Krause, *European Economic Integration and the United States*, Washington 1968.
- R. A. Mundell, "International Trade and Factor Mobility," *Amer. Econ. Rev.*, June 1957, 47, 321-35; reprinted in his *International Economics*, New York 1968, ch. 6.
- S. Pizer and F. Cutler, "Foreign Investments 1964-65," *Surv. Curr. Bus.*, Sept. 1965, 45, 22-32.
- J. D. Richardson, "Theoretical Considerations in the Analysis of Foreign Direct Investment," *Western Econ. J.*, Mar. 1971, 9, 87-96.
- W. S. Salant et al., *The United States Balance of Payments in 1968*, Washington 1963.
- A. E. Scaperlanda and L. J. Mauer, "The Determinants of U.S. Direct Investment in the E.E.C.," *Amer. Econ. Rev.*, Sept. 1969, 59, 558-68.
- A. Schmitz and P. Helmberger, "Factor Mobility and International Trade: The Case of Complementarity," *Amer. Econ. Rev.*, Sept. 1970, 60, 761-67.
- C. Schotta, "The Real Balance Effect in the U.S., 1947-63," *J. Finance*, Dec. 1964, 19, 619-29.
- D. A. Snider, "Capital Controls and U.S. Balance of Payments," *Amer. Econ. Rev.*, June 1964, 54, 346-58.
- E. Spittaller, "A Survey of Recent Quantitative Studies of Long-Term Capital Movements," *I.M.F. Staff Papers*, Mar. 1971, 18, 189-217.
- G. V. G. Stevens, "Fixed Investment Expenditures of Foreign Manufacturing Affiliates of U.S. Firms: Theoretical Models and Empirical Evidence," *Yale Econ. Essays*, spring 1969, 9, 136-93.
- P. Streeten, *Economic Integration: Aspects and Problems*, Leyden 1961.
- H. Theil and A. L. Nagar, "Testing the Independence of Regression Disturbances," *J. Amer. Statist. Ass.*, Dec. 1961, 56, 793-806.
- J. W. Vaupel and J. P. Curhan, *The Making of Multinational Enterprise*, Boston 1969.
- Committee for Economic Development, (CED) *The European Common Market and Its Meaning for the United States*, New York 1959.
- U.S. Department of Commerce, *Balance of Payments Statistical Supplement: Revised Edition*, *Surv. Curr. Bus.* supp. 1963.

# Monopoly Output Under Alternative Spatial Pricing Techniques

By M. L. GREENHUT AND H. OHTA\*

The objective of this paper is to compare the effect on output of spatial price discrimination with that of simple f.o.b. pricing. A possible conclusion would be that if discrimination does not increase output, such pricing is undesirable from the standpoint of social welfare even though it proves to be more profitable for the spatial monopolist. Current subject matter interest thus reflects the question whether the discriminatory delivered price *level* of the spatial monopolist is high or low relative to the set of delivered prices resulting from a policy of f.o.b. mill pricing.<sup>1</sup>

Our problem is, therefore, closely related to the discussion by A. C. Pigou-Joan Robinson and Edgar Edwards of outputs produced by a discriminatory and nondiscriminatory (simple) monopolist. Their formulations of the problem were, unfortunately, too general to provide clear answers to specific questions such as whether or not the existence of economic space helps yield greater (or lesser) outputs under discrimination than does simple monopoly pricing. We shall answer this and related questions precisely by presenting a proof that *monopoly* outputs are *always* greater under spatial price discrimination than they are under simple f.o.b. mill pricing. Throughout the paper we abstract

from special forces, such as adverse income effects associated with discriminatory prices.

## I. A Basic Difference in Nonspatial and Spatial Price Discrimination: Total Output

Classical (nonspatial) economics describes the discriminating monopolist as the supplier of greater *or* smaller outputs than the non-discriminating monopolist, depending on the relative shapes of the demand curves. It can easily be shown, however, that whether greater outputs are produced under discriminatory monopoly also depends in part on the level and/or shape of the cost function. It is this additional property which is the relevant, in fact the basic, force behind the alternative views that apply to nonspatial price discrimination. The purpose of the present section is not to take issue with classical theory, but to reformulate it in a spatial framework from which comparisons may be drawn. Apart from discrimination by licensed practitioners of medicine, law, and dentistry, much of the price discrimination by firms encountered today occurs along geographical lines.

### *The Spaceless Framework*

As a point of departure for our basic model assume, as did Pigou and Robinson, the existence of two markets, each of which possesses a linear demand curve. Let these curves be parallel and differ only in their price and quantity intercepts. Let Market I contain the stronger demand while Market II is characterized by the weaker. Assume that there is no cost of distance or, alternatively, that the economic space separating the markets from the seller is the same.

In Figure 1(a) and (b), the horizontal axes, respectively, measure the individual demand  $q_i$  and the aggregate demand  $Q$  of the two markets. The vertical axes measure the aver-

\* Professor and lecturer at Texas A&M University and Aoyama Gakuin University, respectively. We wish to thank Robert Ekelund and W. D. Maxwell for helpful suggestions and criticisms. This paper is based on research funded by the National Science Foundation.

<sup>1</sup> In another paper, we are examining the impact of economic space on the pricing practices of the spatial monopolist. In contrast to prevailing spatial theory, the proposition is advanced that regardless of the shapes of the assumed (demand) curves, the spatial monopolist increases his profits by absorbing relatively larger quantities of freight on sales to buyers located at greater distances from his plant. The specification that demand vanishes at some high (finite) price along with a resale proviso will be shown to be sufficient to establish the full scope and generality of this proposition.

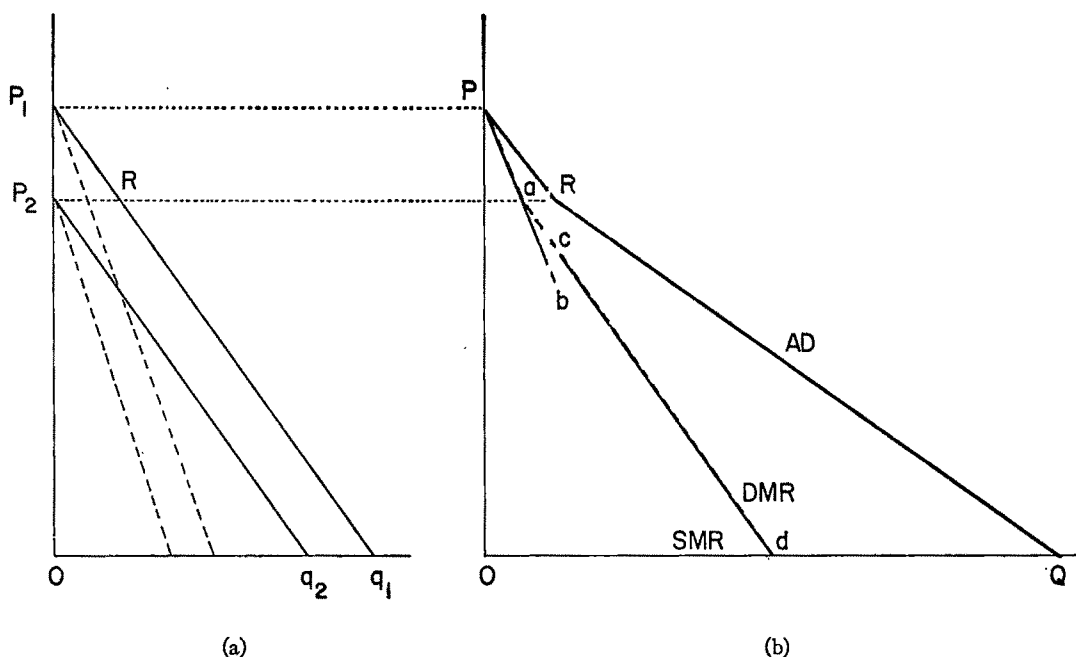


FIGURE 1

age and marginal revenue values for the simple and discriminatory monopolies. The lines  $p_1q_1$  and  $p_2q_2$  in Figure 1(a) stand, respectively, for the average revenue curves, in Market I and Market II. The line  $pRQ$  in Figure 1(b) is the horizontal sum of the two average revenue curves. This aggregate curve ( $AD$ ) applies to a simple monopoly. Its marginal revenue curve,  $SMR$ , is the broken curve  $pbcd$ .<sup>2</sup> Correspondingly, the aggregate demand curve under discriminatory monopoly is also  $AD$ ; but, an average revenue curve is conceptually imaginable which—in light of the different prices charged in the two markets—would yield for the discriminating monopolist a higher net price on sales to both markets at all given-identical total outputs compared to that derived by the simple monopolist. (We do not show this average revenue curve in Figure 1.) The continuous curve  $pacd$  is the dis-

criminatory marginal revenue,  $DMR$ , curve. It is the horizontal sum of the two marginal revenue curves, and must be distinguished from the discontinuous  $SMR$  curve  $pbcd$ . The intersections of  $DMR$  and  $SMR$  with the marginal cost curve determine whether or not total output  $Q$  is greater under discriminatory monopoly than simple monopoly.

**Case I.** If  $MC$  cuts the  $pa$  portion of  $DMR$ , there will be no difference between the outputs produced by a discriminatory monopoly and a simple monopoly. In effect, only one market, namely Market I, is served *even if the monopolist otherwise would practice discrimination*. (See Robinson, p. 196.)

**Case II.** If  $MC$  cuts the  $ab$  portion of  $SMR$ , and also the  $ac$  portion of  $DMR$ , the intersection of  $MC$  with  $DMR$  lies to the right of the intersection of  $MC$  with  $SMR$ . Total output in this case is greater under discriminatory pricing than under simple monopoly pricing. (See Pigou, pp. 286, 809.) Market I alone is served by the simple monopoly whereas both markets are served by the discriminating seller. Correspondingly, the

<sup>2</sup> This diagram, in particular Figure 1(b), is based on Robinson's Figure 65A (p. 201). A diagram by Pigou (p. 809), designed for similar analytical purposes, is inferior to Robinson's in the sense that it requires a constant marginal cost (Pigou's supply price).

equilibrium price is necessarily higher in the case of simple monopoly pricing than  $p_2$  while the price in Market II is lower than  $p_2$  in the price discrimination case. (See Figure 1.)

**Case III.** If  $MC$  cuts the  $cd$  portion of  $DMR$  (which is overlapped by  $SMR$ ), then total output is once again the same under the two different pricing systems. Unlike Case I or II, however, both markets are now served. Although total output remains unchanged, the output shipped to Market II increases by the precise amount by which demand is cut in Market I as a result of the change in the pricing technique from simple to discriminatory pricing.<sup>3</sup>

Of the three cases, Robinson apparently emphasized Case III. She identified it with the rather elementary case of linear demand.<sup>4</sup> Although Case I may be trivial, Case II will be shown to be highly relevant to a spatial analysis of discriminatory pricing.

#### *The Spatial Framework*

To appreciate the relationships relevant to the space economy, conceive of buyers being *evenly distributed* along a line or over a plain with each and every buyer having the same identical (gross) demand curve.<sup>5</sup> Then the net demand curve of a buyer (or a group of buyers) one mile away from the seller must be different (to the seller) than the demand curve of a buyer located at the seller's door. The net demand curve of a buyer two miles away, three miles away, etc. are, in turn, all different in the eyes of the seller.<sup>6</sup> A number

of parallel linear functions can thus be used to portray the net demands confronting the seller. While a single seller cannot serve spatial markets in which the freight cost burden is prohibitive, it should be clear that more distant buyers (or markets) may be served by a discriminating monopolist compared to a simple f.o.b. pricing monopolist. Cases I and III above are, therefore, inapplicable, and Case II alone is relevant under spatial monopoly. But let us now probe into this matter more deeply.

Consider three different net demand curves representing markets located adjacent to, near to, and distant from the seller, respectively. Assume further that: 1) Every buyer's gross demand curve is linear and identical, of the form  $p = b - 3aq$ , and that the aggregate gross demand curve is specifiable as  $p = b - aQ$ , where  $b$  and  $a$  are positive constants. 2) The freight rate (or cost of distance) is zero between the seller and the adjacent market,  $(1/3)b$  over the distance from the seller to the near market, and  $(2/3)b$  for the distance between the seller and the distant market. 3) Each market contains  $\frac{1}{3}$  of the total number of buyers. Figure 2 may, then, be constructed as a simple extension of Figure 1.<sup>7</sup>

Observe that Case III, i.e., where  $DMR$  and  $SMR$  merge into the segment  $cd$  in Figure 1, is now relatively unimportant since the line  $cd$  has become relatively shorter in a threefold market division vis-à-vis the old twofold division of buyers. A new  $df$  portion of  $DMR$  has appeared which is *not overlapped* by  $SMR$ . One might well expect, accordingly, that outputs will often be different.

The following conclusions apply: If mar-

<sup>7</sup> Let  $Q$  stand for the aggregate net demand such that:

$$\begin{aligned} Q &= \frac{b-p}{3a} \forall p \succ p_1 > p \geq p_2, \\ &= \frac{b-p}{3a} + \frac{b-(b/3)-p}{3a} \forall p \succ p_2 > p \geq p_3, \\ &= \frac{b-p}{3a} + \frac{b-(b/3)-p}{3a} + \frac{b-(2b/3)-p}{3a} \\ &\forall p \succ p_3 > p \geq 0 \end{aligned}$$

This aggregation provides the kinked demand curve  $pRSQ$ . Similarly, the  $DMR$  and  $SMR$  curves are specifiable rigorously, as given later.

<sup>3</sup> See Pigou, p. 809; Robinson, p. 192. And see R. Battalio and Ekelund for a discussion of the importance of cost in output changes under price discrimination.

<sup>4</sup> Compare Robinson, p. 192. More generally, she observed that output is the same when the relative adjusted concavities of the two demand curves are the same.

<sup>5</sup> This spatial system conforms to the framework used by Edgar Hoover, Arthur Smithies, and Donald Dewey in which buyers are evenly distributed along a line or over a plain, individual gross demands are identical, and freight rates are significantly positive.

<sup>6</sup> In other words, the spatial gross demand function, i.e., the demand function at the customer's own mill, must be distinguished from its related net demand function, i.e., the demand visualized by the seller as applicable *at the seller's site*. The two are sharply distinct when freight costs are significant.

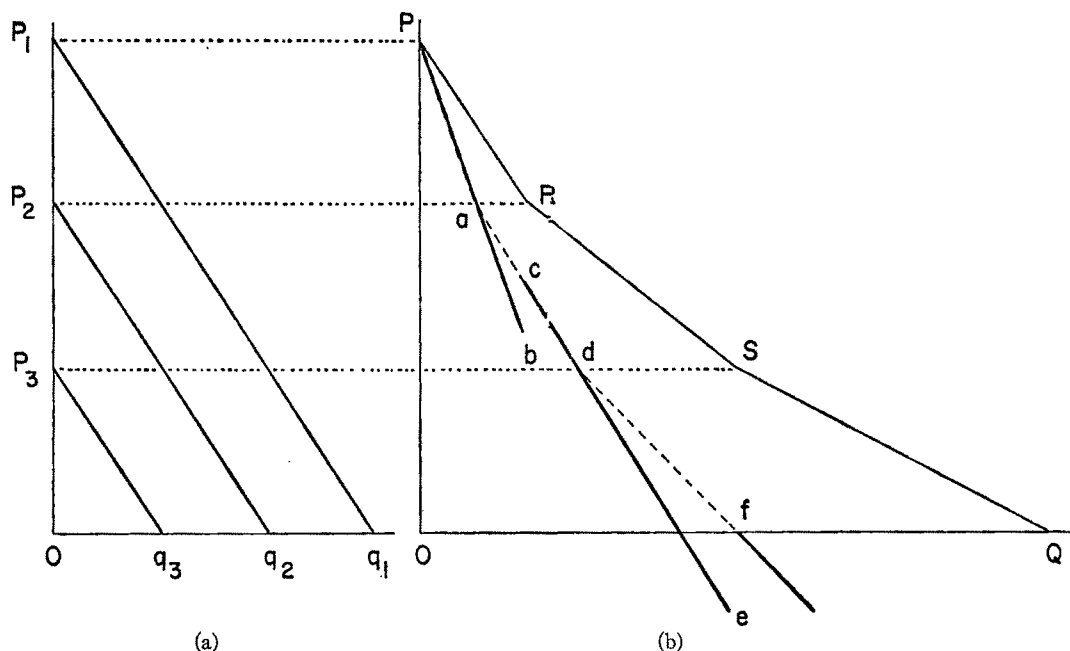


FIGURE 2

ginal cost is so high that the  $MC$  curve cuts the  $ac$  portion of  $DMR$ , total output would clearly be greater under discrimination than it would be under simple monopoly. If the  $MC$  curve cuts the  $cd$  portion of  $DMR = SMR$ , the total outputs would be the same. If  $MC$  were still lower so that it cuts the  $df$  portion of  $DMR$ , total output once again would be greater under discriminatory pricing than simple monopoly pricing. Diverging total outputs become more and more likely under linear demand, the "finer" the division of markets in economic space.

## II. Spatial Equilibrium Given The Linear Gross Demand Curve

Consider the "line" situation where  $n$  spatially separated markets of the same size are identifiable, and each buyer's gross demand curve is linear of the form  $p = b - naq$ . The freight rate per unit of distance may now be defined by the constant value  $b/n$ , where  $b$  stands for the price intercept of the gross demand curve of a buyer at a given site on the line market. Thus, for example, the buying market one distance unit from the seller

involves a freight rate equal to  $b/n$ , the buying market two distance units from the seller involves a freight rate equal to  $2b/n$ , etc. The price intercepts ( $p_1, p_2, \dots, p_n$ ) of the net linear demand curves are, therefore, definable as  $p_i = b - (i-1)b/n$ , where  $i = 1, 2, \dots, n$ . (See Figure 3.) The aggregate demand function may then be established arithmetically by summing all of the net demand curves. We obtain (1), as derived below,<sup>8</sup> and (1') similarly.<sup>9</sup>

<sup>8</sup> With respect to (1), consider the following:

$$\begin{aligned}
 Q &= \frac{b-p}{na} \quad \forall p \succ p_1 > p \geq p_2, \\
 &= \frac{b-p}{na} + \frac{b-(b/n)-p}{na} = \frac{2b-(b/n)-2p}{na} \\
 &\quad \forall p \succ p_2 > p \geq p_3, \\
 &= \frac{2b-(b/n)-2p}{na} + \frac{b-(2b/n)-p}{na} \\
 &= \frac{3b-(3b/n)-3p}{na} \quad \forall p \succ p_3 > p \geq p_4, \\
 &= \frac{3b-(3b/n)-3p}{na} + \frac{b-(3b/n)-p}{na}
 \end{aligned}$$

$$(1) \quad Q = \frac{i[(b-p) - (i-1)b/2n]}{na}$$

$\forall p \nmid p_i > p \geq p_{i+1} \quad (i = 1, 2, \dots, n),$   
and

$$(1') \quad \lim_{n \rightarrow \infty} Q = \frac{(b-p)^2}{2ab} \quad \forall p \nmid b \geq p \geq 0$$

The discriminatory marginal revenue, *DMR*, is obtainable via the horizontal summation of all net individual marginal revenue curves, with each horizontal intercept value being one-half that of the respective net demand curve. Dividing the right-hand side of (1) by two, therefore, yields *DMR* or, more rigorously, it yields the inverse function of aggregate marginal revenue under price discrimination. Thus:

$$(2) \quad Q = \frac{i[(b-p) - (i-1)b/2n]}{2na}$$

$\forall p \nmid p_i > p \geq p_{i+1} \quad (i = 1, 2, \dots, n),$   
and

$$\begin{aligned} &= \frac{4b - (6b/n) - 4p}{na} \quad \forall p \nmid p_4 > p \geq p_6, \\ &= \frac{i(b-p) - i(i-1)b/2n}{na} \quad \forall p \nmid p_i > p \geq p_{i+1}, \end{aligned}$$

where  $p_i = b - (i-1)b/n$ ,  $i = 1, 2, \dots, n$ , and  $n$  stands for the number of submarkets *visualized*, but not necessarily served by the seller.

<sup>9</sup> As  $n$  approaches a large number, any  $p$  in the domain  $p_i > p \geq p_{i+1}$  can be approximated by:

$$\begin{aligned} p &= (p_i + p_{i+1})/2 \\ &= [2 - (2i-1)/n](b/2), \end{aligned}$$

by substitution of price intercepts  $p_i$ ;

$$\therefore i = (1/2) + (b-p)n/b$$

In order to define  $Q$  in terms of  $p$  alone, substitute this value of  $i$  into the general formula for  $Q$ , which yields:

$$\begin{aligned} Q &= \frac{b-p}{2na} - \frac{b[(b-p)(n/b) - (1/2)]}{4n^2a} \\ &\quad + \frac{(b-p)^2}{ab} - \frac{(b-p)^2(n/b) - (b-p)/2}{2na}, \end{aligned}$$

$$\lim_{n \rightarrow \infty} Q = \frac{(b-p)^2}{2ab} \quad \forall p \nmid 0 \leq p \leq b$$

$$(2') \quad \lim_{n \rightarrow \infty} Q = \frac{(b-p)^2}{4ab} \quad \forall p \nmid b \geq p \geq 0,$$

where  $p$  now stands for values to be read along the marginal revenue curve. This formula (2), or (2'), would provide the total output produced by the seller. Moreover, the formulas permit comparison of spatial output with the spaceless economy output. Thus, if marginal cost and marginal revenue were zero,  $p$  in (2) assumes the particular value  $p = p_{n+1}$ . Substituting  $i = n$  and then  $p = 0$  in (2) and (2'), respectively, yield:

$$(2^*) \quad Q = \frac{(n+1)b}{4na},$$

and

$$(2^{*'}) \quad \lim_{n \rightarrow \infty} Q = \frac{b}{4a}$$

Comparison of (2<sup>\*</sup>) with (2<sup>\*</sup>) at  $p = 0$  indicates that when marginal production cost is zero, the spatial output  $Q = b/4a$  is less than the output  $Q = b/2a$  of the spaceless ( $n = 1$ ) economy. In other words, if buyers are successively (evenly) distributed along a line, total output under discriminatory monopoly would, in the limit, be reduced to half that which would have been produced if all buyers had been concentrated at sites next to the seller.<sup>10</sup>

<sup>10</sup> The ratio  $r$  of spatial demand to spaceless demand applicable at any marginal cost = marginal revenue level is given by equation (2) divided by the counterpart spaceless marginal revenue equation  $Q = (b-p)/2a$ . Thus:

$$\begin{aligned} (a) \quad r &= \left( \frac{i[b-p - (i-1)b/2n]}{2na} \right) / \left( \frac{b-p}{2a} \right) \\ &= \frac{i[b-p - (i-1)b/2n]}{(b-p)n}, \end{aligned}$$

where  $i$  is an integer whose value depends on  $p$  such that  $p_k > p \geq p_{k+1} \Rightarrow i = k$ ; e.g., if  $p_3 > p \geq p_4$ , then  $i = 3$ . This formula (a) reappears in much simpler form under certain conditions. To derive this simpler form, it helps to recall from the definition  $p_i = b - (i-1)b/n$  that the price intercept of the strongest market is given by  $p = b$ ; in turn, the price intercepts of the other spatial markets are fully specified by  $k+1$ , where  $i = k = 1, 2, \dots, n$ .  
(over)

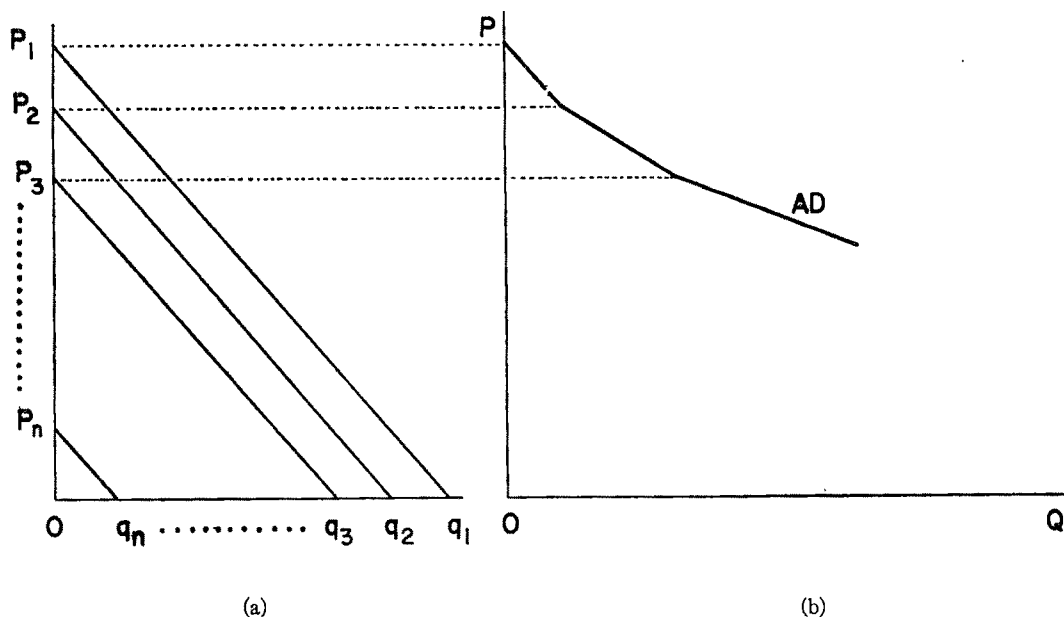


FIGURE 3

Consider now the form of *SMR* for later comparison with *DMR*. Observe first that it is *not* a continuous function *nor* a one-to-one function. (See Figure 1 or 2.) In other words, even though  $p$  may be taken as a function of  $Q$ , the  $Q$  relating to *SMR* is not a function of  $p$  since more than one  $Q$  value relates to some of these  $p$ 's. Since *SMR* is not a one-to-one function, no inverse function may be derived for it that would compare with (2). We may, nevertheless, rewrite (2) as (3) and subsequently specify a new domain for it:

$$(3) \quad p = b - \frac{(i-1)}{2} \frac{b}{n} - \frac{2na}{i} Q \quad (i = 1, 2, \dots, n),$$

where, to repeat, the domain must be re-defined to establish a different function. We define this domain in (4), as derived in turn below:<sup>11</sup>

$$(4) \quad \frac{(i+1)ib}{2n^2a} > Q \geq \frac{(i-1)ib}{2n^2a} \quad (i = 1, 2, \dots, n)$$

If then *MC* is at the level  $p_{k+1}$ , both  $i$  and  $p$  in equation (a) have been provided; viz.,  $i=k$  and  $p=p_{k+1}=b-kb/n$ . Substituting these particular values of  $i$  and  $p$  in (a) yields the desired relation:

$$(b) \quad r = \frac{k+1}{2n}, \quad k = 1, 2, \dots, n$$

Formula (b) is not defined when  $k=0$ , since this condition implies  $p=b$  in (a), and that the denominator (as well as numerator) of formula (a) are zero. Most importantly, formula (b) demonstrates that the lower is *MC* ( $=MR=p$ ), i.e., the larger is  $k$ , the higher is the ratio of spatial to spaceless demand. *Ceteris paribus*, when *MC*=0 the ratio 1/2 applies, which indicates that in the limit aggregate spatial demand is one-half that of spaceless demand.

<sup>11</sup> The domain  $Q$  for *SMR* may be partitioned for each point of  $Q$  at which a kink in the aggregate demand *AD* occurs. Following previous definitions, and assumptions of freight rates and market distances, the price intercepts ( $p_1, p_2, \dots, p_n$ ) of the net linear demand curves occur at the values  $p=p_i=b-(i-1)b/n$ ,  $i=1, 2, \dots, n$ . Substituting for  $p$  in (1) then yields the critical values of  $Q$  at which kinks of *AD* occur, namely:

$$Q_i = \frac{i(i-1)b}{2n^2a}$$

and

$$Q_{i+1} = \frac{(i+1)ib}{2n^2a} \quad (i = 1, 2, \dots, n)$$

The relevant partitioned domain is defined above as (4).

Formulas (4) and (3) establish the domain and the corresponding range of the marginal revenue of the simple (nondiscriminatory) monopolist. From elementary economics, we know that the level of  $p$  in (3) is determined by the level of  $MC$ . Hence the total output produced under simple monopoly, given the  $MC$ , can be obtained by (3) provided that  $i$  is also given. And  $i$  (the number of markets served) turns out to be a function of  $p$  as  $n$  approaches a large number. By elementary substitution,<sup>12</sup> we obtain:

$$(5) \quad \frac{i}{n} = \frac{2(b-p)}{3b}$$

Substituting (5) into (3) and taking the limit yields:

$$(6) \quad \lim_{n \rightarrow \infty} Q = \frac{2(b-p)^2}{9ab}, \quad \forall p \geq b \geq p \geq 0$$

The single formula (6) provides the total output produced under simple spatial monopoly as  $n$  approaches infinity.<sup>13</sup> We see

<sup>12</sup> Substitute the inverse of (3), i.e., (2), into (4). This process yields:

$$(a') \quad \frac{2n}{3b}(b-p) + 1 \geq i > \frac{2n}{3b}(b-p) - \frac{1}{3}$$

Because  $i$  is an integer and the interval of  $i$  is greater than unity, i.e.,  $4/3$ , inequality (a') provides either one or two values for  $i$  depending on the value of  $p$ . Thus,  $i$  in general is not a function of  $p$ . This result stems, of course, from the fact that  $SMR$  is not a one-to-one function. Nevertheless, (a') can be rewritten in the more revealing form (b'):

$$(b') \quad \frac{2}{3b}(b-p) + \frac{1}{n} \geq \frac{i}{n} > \frac{2}{3b}(b-p) - \frac{1}{3n}$$

This indicates that if  $n$  is a large number,  $i/n$  may be approximated by  $2(b-p)/3b$ , as in (5).

<sup>13</sup> It might also be noted that (6) conforms to the limiting aggregate demand function (1'), actually being its marginal curve. To see this, consider the inverse of (1'), i.e., (1'') below:

$$(1'') \quad p = b - \sqrt{2abQ^*}, \quad \text{where } Q^* = \lim_{n \rightarrow \infty} Q$$

The total revenue and the marginal revenue are respectively given by:

$$(a'') \quad pQ^* = bQ^* - (2abQ^*)^{1/2}Q^*,$$

$$(b'') \quad \frac{d(pQ^*)}{dQ^*} = p = b - \frac{3\sqrt{abQ^*}}{\sqrt{2}};$$

that when the spatial market is "finely" divided,  $Q$  becomes a one-to-one function of  $p$  over the domain  $b \geq p \geq 0$ . This fine division also implies that the marginal revenue curve under simple monopoly, i.e.,  $SMR$ , approaches a continuous curve, as shown in Figure 4. An alternative (direct) proof of this continuity relation simply requires substitution of the lower limit value (i.e., the smallest output value  $Q$  when  $i=k$ ) and then the upper limit value (i.e., the largest output value  $Q$  when  $i=k-1$ ) of (4) into (3). Doing so yields:

$$(7) \quad p_i = b - \frac{(i-1)}{2} \frac{b}{n} - \frac{2na}{i} \frac{i(i-1)b}{2n^2a} \\ = b - \frac{3ib - 3b}{2n},$$

and

$$(7') \quad p'_i = b - \frac{(i-2)b}{2n} - \frac{2na}{i-1} \frac{i(i-1)b}{2n^2a} \\ = b - \frac{3ib - 2b}{2n},$$

where (7) gives the upper value for  $p$  at the kink in the aggregate demand curve and (7') gives the lower value for  $p$ .<sup>14</sup> It follows that  $p_i - p'_i = b/2n$ , and:

$$(8) \quad \lim_{n \rightarrow \infty} (p_i - p'_i) = 0$$

Compare now the limiting  $SMR$  in (6) with the limiting  $DMR$  which was derived in (2'). Remarkably, the value  $Q$  in (2') is

$$(c'') \quad Q^* = \frac{2(b-p)^2}{9ab}$$

<sup>14</sup> For a specific example, consider the limiting value  $p$  in equation (3), given  $i=3$ . We obtain  $p = b - b/n - (2/3)naQ = (n-1/n)b - (2/3)naQ$ . Next, inserting the same  $i=3$  in (4) yields  $((4)3b/2n^2a) > Q \geq 3(2)b/2n^2a$ . The upper value for  $p$  at the kink in the aggregate demand curve is then derived by substituting in  $p = (n-1/n)b - (2/3)naQ$  the lower limit value of  $Q$  when  $i=3$ , i.e., the smallest output of  $Q$ , as in (7). In turn, the lower value of  $p$  at the kink under simple monopoly is given by substituting into (3) the upper limit value of  $Q$  when  $i=2$ , i.e., the greatest output, as in (7').

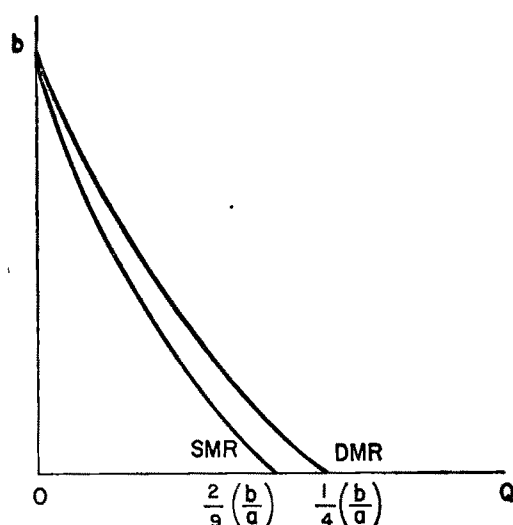


FIGURE 4

strictly greater than  $Q$  in (6) for all  $p$  such that  $b > p \geq 0$ , as is shown below:

$$(9) \quad \frac{(b-p)^2}{4ab} > \frac{2(b-p)^2}{9ab} \quad \forall p \mid b > p \geq 0$$

This relation is sufficient to justify Figure 4. Besides continuity, we see in Figure 4 that the  $DMR$  curve always lies above the  $SMR$  curve, except at  $p=b$ . Manifestly, the quantity produced in the case of linear gross demand must always be greater under spatial monopoly price discrimination than under simple f.o.b. spatial monopoly pricing.

### III. The Spatial Model Generalized for Non-Linear Demand

Our thesis that total output increases under spatial price discrimination holds true regardless of the shape and/or level of the  $MC$  curve. Moreover, it does not depend on the linear demand curve assumption, since non-linear curves may be approximated by linear lines over small intervals. To obtain aggregate demand, minuscule linear portions of any (concave or convex) demand curve may be added successively. The resulting  $SMR$  and  $DMR$  curves must then have fundamentally the same characteristics as those specified above at the conclusion of Section

II of this paper. This basic idea supports our proposition that total output is greater under *spatial* discrimination than under simple spatial monopoly *regardless of the shape of the gross demand curve*.<sup>15</sup>

### IV. Conclusions

Spatial price discrimination, we have therefore shown, yields larger outputs than does simple f.o.b. mill pricing, *regardless of the shape of the gross demand curve*. This conclusion justifies Pigou's anticipation that total output is likely to be greater under discriminatory monopoly than simple monopoly as the number of separate markets increases (see p. 287). But Pigou based his expectation on the claim that third degree discrimination blends into first degree discrimination as the number of separate markets increases, an identity requiring a perfectly inelastic demand curve in each sub-market and, in turn, complete appropriation of consumer surplus.<sup>16</sup> Pigou's expectation of greater outputs was, therefore, circumscribed unnecessarily. From a somewhat different perspective, Robinson also stressed output effects.<sup>17</sup> She pointed, in general, to the in-

<sup>15</sup> Direct proof of this proposition will be provided to interested readers upon request of the authors.

<sup>16</sup> See Pigou, p. 287, where he states that third degree discrimination approximates first degree discrimination as the number of markets increases. To recall his formulations, the first degree involves the charging of a different price for each unit of the commodity sold such that the price for each unit exacted equals its demand price; that is to say, no consumer's surplus at all is left to any buyer. The second degree of discrimination is simply an incomplete form of the first degree; it involves a different price for a different set of units of the commodity with some consumer's surplus left over. The third degree would obtain if the monopolist were able to classify his customers by groups . . . and could charge a different price to the members of each group, as in Figure 1.

<sup>17</sup> Robinson argued contra Pigou that whether or not output is greater under discrimination than simple monopoly ". . . depends not upon the number of markets but upon the relative concavities of the separate demand curves" (p. 201). More specifically, she claimed that total output would be greater or less under monopoly price discrimination depending upon whether the more elastic of the demand curves in the separate markets is more or less concave than the less elastic demand curve; in turn, outputs would be the same if the demand curves are straight lines or if the relative adjusted

determinacy of total output quantities under simple monopoly and discriminatory pricing.<sup>18</sup> Our own basically general formulation applies to economic space, and relates to the inevitable separation of *DMR* and *SMR*.

It is useful at this point to recall a by-product of our analysis, namely that total demand is less when buyers are divisible into  $n$  different submarkets along a line than when they are all concentrated at the seller's location. An even more compelling view of the distinction between spatial and non-spatial output is indicated by the rule that the output of the former compared to the latter is less the higher is the level of marginal costs. It follows that if production costs and freight costs are significant, simple aggregation of gross demand curves must involve serious overestimation of effective demand. Identical gross demands over space are, in fact, significantly different net demands when the friction of distance is great.

The several conclusions drawn above stem from the assumptions that buyers are evenly distributed along a line or over a plain with each having identical tastes and hence identical gross demand curves. These assumptions are, however, neither intrinsic nor crucial to our basic conclusions. Just so long

concavities are equal. Her argument holds only if *all markets* are served under simple monopoly as well as under discriminatory monopoly. And see Robinson, pp. 190-194. In other words, her argument is based on Case III in Section I.

<sup>18</sup> Also see F. M. Scherer where, pursuant to Pigou's formulation, he concludes that if demand functions are linear, "... output under discrimination will be identical to output under simple uniform-price monopoly ...", and where pursuant to Robinson's analysis, he states "... it is impossible to determine whether on balance third degree discrimination increases output ..." (p. 254).

as successively shrinking net demand curves characterize the demands of more distant buyers, our conclusions hold regardless of other facets of buyer distribution and/or tastes.<sup>19</sup> In the space economy, price discrimination always yields greater output for the spatial monopolist than does simple f.o.b. pricing.

# REFERENCES

- R. Battalio and R. B. Ekelund, Jr., "Output Change Under Third Degree Discrimination," unpublished manuscript.
- D. Dewey, "A Reappraisal of F.O.B. Pricing and Freight Absorption," *Southern Econ. J.*, July 1955, 22, 48-54.
- E. O. Edwards, "The Analysis of Output under Discrimination," *Econometrica*, Apr. 1950, 18, 168-72.
- M. L. Greenhut, *A Theory of the Firm in Economic Space*, New York 1970.
- and H. Ohta, "The Classical Theory of Spatial Price Discrimination," unpublished manuscript.
- E. M. Hoover, "Spatial Price Discrimination," *Rev. Econ. Stud.*, June 1937, 4, 182-91.
- A. C. Pigou, *The Economics of Welfare*, 3d ed., London 1929.
- J. Robinson, *The Economics of Imperfect Competition*, London 1933.
- F. M. Scherer, *Industrial Market Structure and Economic Performance*, Chicago 1970.
- A. Smithies, "Monopolistic Price Policy in a Spatial Market," *Econometrica*, Jan. 1941, 9, 63-73.

<sup>19</sup> What we have shown in the present paper is that the Case III requirement cannot apply if the monopolist faces innumerable many weaker and weaker markets along with a strongest market, the situation inherent to a space economy, *ceteris paribus*. See Greenhut, pp. 113-14.

# Job Search, the Duration of Unemployment, and the Phillips Curve: Comment

By PAUL GAYER AND ROBERT S. GOLDFARB\*

Dale Mortensen's paper in a recent issue of this *Review* is a valuable contribution to the growing literature relating job search to Phillips curves. This note argues that a broadening of the definitions of uncertainty used in the Mortensen article—and many of the earlier job search-Phillips curve articles<sup>1</sup>—would result in more fruitful models. Specifically, the general result obtained by these articles—that there is no *long-term* tradeoff between inflation and unemployment—might be changed.

The job search literature allows uncertainty on the part of workers as to which employers are paying which wages. Mortensen himself also allows different employers to require different skill levels for vacancies, so that workers are also uncertain about skill requirements attached to particular vacancies:

The only information possessed by the participant is the nature of the frequency distribution of all offers and the structure of wage offers across skill levels. To determine whether his own skills meet the requirements for a particular vacancy and to ascertain the wage offered to fill the opening, he must search it. [p. 848]

These assumptions lead to positive unemployment in equilibrium in a labor market with a constant flow of new entrants; the unemployment represents the time it takes for new entrants to find acceptable jobs.

In all of this, there is no uncertainty on

\* Staff member, Office of Planning, Research and Evaluation, Office of Economic Opportunity, and assistant professor of economics, Yale University, respectively. We wish to thank Michael Barth, Daniel Hamermesh, Jim Hosek, and Alvin Klevorick for helpful comments. Views expressed in the paper are solely the authors'.

<sup>1</sup> A number of these articles, and a review of the literature by Edmund S. Phelps, appear in the volume edited by Phelps.

the employer's side about the quality of labor. The firm sets a minimum qualifications level by making the appropriate cost minimization calculation, and has no trouble judging whether a particular individual who applies meets this qualification or not. *These qualifications, it should be stressed, are accurate proxies for the applicant's true marginal productivity:* the firm can evaluate an individual applicant's productivity quite precisely.

We would propose to Mortensen and other trustees of the job search-Phillips curve school that they allow for employer uncertainty in evaluating applicants. This would have two advantages. First, it would increase the "realism" of search models. There is abundant evidence that the employer himself faces very serious difficulties in evaluating the quality of applicants, and uses a number of peculiar devices to deal with the problem. Second, the incorporation of employer uncertainty changes the implication of Mortensen's and other search models that the Phillips curve tradeoff is only transitory.

Let us first defend the assertion that the employer faces serious difficulties in evaluating applicants. There are many human qualities which contribute to employee productivity; many of the most important of these qualities are not observable or testable by the employer in the short period of time over which he must decide whether or not to offer a job. Consider, for example, a job which only involves "pushing a button"; anyone with fingers can do it. The productivity of the button-pusher will depend on how reliable he is about showing up every day on time, whether he will be willing to work extra hours when there is an emergency, and so forth. It is not Mortensen's skill requirements which completely determine his productivity.

Employers are well aware of the difficulty of evaluating real productivity quickly. To

deal with this complex search problem, several devices are used. A major device is the use of "hiring standards" or "screening devices" to pick out those applicants most likely to have such nonobservables as motivation and reliability. Given a choice between applicants with high school diplomas and those without, the manufacturer hiring semi-skilled production workers will often choose those with high school diplomas; the hiring standard or screening device is a high school diploma. Even if the job just involves—as it frequently does—button pushing, the individual with the high school diploma is more likely to be productive. He was motivated enough and reliable enough to stick it out through high school, so that he is likely to be motivated enough and reliable enough to show up every day at the job. All kinds of screening devices which are hard to explain when viewed in terms of the specific technical skill requirements of the job (button pushing in our example) become easy to understand when the view of a productive worker is broadened. A recent study of the Chicago labor market by Albert Rees documents the variety of hiring standards and screening devices used:

An applicant for employment may be examined in several or in extreme cases all of the following ways: a written application for employment, an interview, paper and pencil tests, work sample tests, a medical examination, a check of credit standing, a check of school and employment references, and even police record checks . . . we encounter such rules as the following: clerical workers must be high school graduates; material handlers must weigh at least 150 pounds; janitors must have lived a year in the metropolitan area; employees who use public transportation must not need to make more than two transfers. . . . In addition to formal hiring standards, employers have a still more flexible set of preferences among job applicants, such as the preference for married men for unskilled work because they are thought to have lower quit rates.  
[pp. 561–62]

The existence of this quality evaluation

problem on the employer side would seem to have important implications for the unemployment-inflation tradeoff. At a given point in time, the employer knows with relative certainty the productivity of those employees who have been in his employ for a reasonable length of time. He does not know the productivity of applicants. All he can do is make educated guesses by using hiring standards to pick out those candidates who seem most likely to produce at or above some expected level.

At a given level of aggregate demand, the existence of hiring standards implies that there may be some individuals unemployed not merely because they are searching (that is, looking for a "high enough" wage), but because they fail to meet the minimum hiring standards set by the most unselective firm in the economy. No employer is willing to hire these individuals.<sup>2</sup>

Why don't wages drop enough to ensure the employment of all these rejected workers? Institutional or organizational constraints might include legal minimum wage levels, laws against race, sex and age discrimination, and entry wage levels set by unions. Further, it is at least possible that some rejects have low enough expected marginal products that no positive wage can be paid.<sup>3</sup>

Now suppose that aggregate demand rises, so that all firms experience a rise in product prices. As they attempt to expand employment at the same or higher wages, some firms discover that they can do so only by lowering hiring standards, hiring some previous rejects.

So far the story is no different from Mortensen's—he too expects a fall in skill requirements at particular firms when labor markets tighten. The crucial addition which

<sup>2</sup> Notice that there is no explicit minimum level of acceptable skills in Mortensen's model.

<sup>3</sup> This assumes that there are some fixed hiring costs per worker, so that the worker's marginal product must be enough above the wage to recoup these hiring costs over the expected stay of the employee. With low expected marginal productivity and high probability of the employee quitting quickly, it is possible that no positive wage exists which creates a large enough gap between marginal product and wage to make hiring the individual profitable.

our approach allows arises from the information gains about worker quality which accrue to employers as they try out those workers who previously failed to meet previous hiring standards. Some of these workers will turn out to have marginal products higher than that expected on average by the employer when he set his hiring standard. Some people without high school diplomas are just as motivated and productive as those with high school diplomas. In Mortensen's world these information gains from inflation do not exist, because employers are omniscient about the quality of labor of each applicant; in our world, these gains do exist because the employer is able to better specify the productivity of particular hired applicants by seeing them perform on the job—our world allows “learning by hiring.”

How does this increased information influence the inflation-unemployment tradeoff? Let us compare two economies with identical labor forces and identical formal hiring standards.

Although the two economies now have zero rates of price increase, one of them previously experienced rapid inflation over a long period, the other economy did not. Since employers in the formerly inflationary economy have more information about the productivity of workers—in particular, they now can identify high productivity workers not meeting formal hiring standards—employment will be higher in the economy. *People who, in the noninflationary economy, would be viewed as unemployable, as having too low an expected productivity, would be employable in the economy with the inflationary history. That is, the long-run equilibrium rate of unemployment in the formerly inflationary economy would be lower.*<sup>4</sup>

<sup>4</sup> The argument that history matters is actually extendable in several directions. Above we argued that the same employer would be likely to retain some people not meeting “after inflation” hiring standards. One can also argue that the worker who is let go after the inflation may have a better chance of being hired by other employers than he did before the inflation. This may be true for two reasons: first, his work experience during the inflation gave him some skills which increase his expected marginal product to other employers, and, second, that *the work experience itself* may make him meet certain hiring standards which he did not meet

The fundamental reason for this difference in implications between Mortensen and us is clear; in Mortensen, unemployment only exists because individuals have acceptance wages higher than the offers they can get from employers. There are always employers who will hire them at some low enough wage. *In our world, there are people who cannot get hired no matter how much they lower their acceptance wage. Further, some of these “unhirables” are not really unhirable: they only look unhirable to the imperfectly informed employer. Their actual marginal product is higher than the marginal product the employer expects them to have. Inflation forces the employer to try out some unhirables; this results in his obtaining a better estimate of the unhirables’ true productivity. This gain in information results in higher employment after the inflation. An economy which was systematically underestimating the productivity of some labor will employ more labor when this systematic underestimation is removed.*<sup>5</sup>

Our argument has involved gains in information about the quality of labor. Similar results might be derived from assumptions that inflation produced better information about the production function facing the firm. If an inflationary period causes the firm to operate on a part of its production function it had not operated on before, the firm might discover that it had overestimated actual production costs at this new larger

---

before. This can happen when the hiring standard is in terms of previous employment experience. Thus, an individual who held a job for eight months and then was laid off may be more attractive to an employer than one who has held no job recently.

<sup>5</sup> The reader may wonder why we assume systematic underestimation; why isn't systematic overestimation also a possibility? There are two answers: in the first place, the employer's hiring standard may be correct on average. That is, the mean performance of people with less than eight years of education (for example) will be unacceptably low. But the population of people with eighth grade educations contains some variance; some people perform better than the average. Second, it would not be surprising if risk-averting employers maintained hiring standards a bit higher than “necessary” (i.e. corrected the hiring standard by a risk premium); it would be extremely surprising if they maintained over time hiring standards that were “too low.”

output and lower average quality of labor point. This type of learning could result in higher employment after the inflation.

## REFERENCES

D. Mortensen, "Job Search, The Duration of Unemployment, and the Phillips Curve,"

*Amer. Econ. Rev.*, Dec. 1970, 60, 847-62.

E. S. Phelps, *Microeconomic Foundations of Employment and Inflation Theory*, New York 1970.

A. Rees, "Information Networks in Labor Markets," *Amer. Econ. Rev. Proc.*, May 1966, 56, 559-66.

# Job Search, the Duration of Unemployment, and the Phillips Curve: Reply

By DALE T. MORTENSEN\*

Paul Gayer and Robert Goldfarb (G-G) are certainly correct when they claim that job qualifications are imperfect predictors, not accurate measures of the potential productivity of prospective workers. Their statement that my neglect of this distinction and the learning asymmetry, namely that employers have the opportunity to learn about the true productivity of any employee but not about the productivity of any rejected applicant, are responsible for my conclusion that there is no *long-term* tradeoff between unemployment and inflation may also be correct. However, G-G confuse the issue by appealing for the relevance of "institutional and organizational constraints" on wages; this argument is not needed to make their case. In my opinion, they have also overlooked potentially important contributions to our understanding of "hidden unemployment," "underemployment," and aggregate labor productivity. In this reply I attempt to clarify their argument on the Phillips curve issue and to point out some of these other contributions.

To see clearly the implications of the G-G hypothesis for the Phillips curve, we must contrast the equilibrium situation implied by the model analyzed in my paper with that implied by the model as they amended it. For the purpose of the comparison, *we explicitly assume that there are no artificial constraints on the eventual adjustment of wages.* Hence, in either situation, the wage structure defined as a relationship between wage offers and qualifications will adapt so that all of those willing to work, given the structure, will be able to after an appropriate period of search.

This conclusion appears to be at odds with the G-G statement, elaborated in footnote 3, that wages at the lower end of the structure

may have to be negative. This appearance of conflict is resolved when one recognizes the fact that participants will decide not to supply their services well before the wage offered falls to zero. In terms of classical supply and demand analysis, it is reasonable to expect, for the reasons given by G-G, that the demand curve for labor whose qualifications are sufficiently poor will intersect the wage axis at a level equal to no less than the wage at which the supply curve intersects the same axis. In other words, no worker with these qualifications is willing to work in return for either his *actual* net marginal product in my world or his *expected* net marginal product in the world of Gayer and Goldfarb.

Although the characteristics of equilibrium in the two cases can be described in identical terms, the interpretations differ significantly. First, consider the case in which qualifications and productivity are equivalent. Although we may lament for those whose low productivity have forced them to decide to stay home with the kids or engage in some other nonmarket activity and we may recommend training programs and minimum incomes as solutions to their economic plight, we cannot objectively classify them as involuntarily unemployed. However, if qualifications are only related to productivity in a stochastic sense, then there will be some among those who have chosen not to work who are nevertheless productive enough to earn a wage at which they would be willing to work if the true facts were known to the appropriate employer. Their unemployment is "hidden" behind a veil of imperfect information and will remain so because the existing information exchange mechanisms will not reveal the truth in equilibrium. One is hard pressed to classify the situation of such workers as anything other than involuntary unemployment.

What about the Phillips curve? Having

\* Associate professor, Northwestern University.

conceded that a pool of willing and able workers without employment can exist, even in the absence of artificial restraints on wages, and now conceding that some of these will be drawn into the labor force during an expansion and will be retained even after expectations have fully adjusted for the reasons given by Gayer and Goldfarb, I must also admit that the long-run Phillips curve may not be vertical. A number of qualifying points must be made, however. First, these workers will not be counted as unemployed in the official statistics of any country for which empirical Phillips curves have been fitted. Hence, the hypothesis cannot be expected to account for the failure of any test of the vertical long-run slope of any such curve. Second, the increased employment attributable to the information gained about the productivity of poorly qualified workers during any single expansion will deteriorate at the attrition rate. An expansion is needed to screen each new generation of workers. Third, alternative information-exchange mechanisms designed specifically to provide the correct information may be cheaper from a social point of view than a sequence of induced inflations.

In a world which offers heterogeneous jobs to a heterogeneous labor force the uncertainty in the relationship between qualifications and productivity and the learning asymmetry may also be responsible in large

measure for "underemployment." In equilibrium some workers will be employed in jobs which require less than their full capabilities because their qualifications underestimate their true talents. Given that employers are willing to reduce qualifications in periods when the labor market is tight, as my model suggests, an expansion offers such workers the opportunity to move to jobs with responsibilities and wages more in line with their capacities. This fact and the fact that an expansion results in the initial employment of some workers whose qualifications overestimate their productivity have implications for the behavior of average labor productivity during the expansion phase of the business cycle.

Once qualifications are lowered, average productivity may fall sharply. However, as those employed whose qualifications overestimate their productivity are discovered, a process which will be distributed over time, average productivity rises again. Hence, the author's hypothesis seems to offer a rather simple explanation for what have been puzzling facts.

Obviously, I feel that the contribution of Gayer and Goldfarb could shed needed new light on a number of important issues. Hopefully, my comments will encourage them to provide us with a more complete and rigorous analysis of their hypothesis and its implications.

# Production Indeterminacy with Three Goods and Two Factors: Rejoinder

By DOUGLAS B. STEWART\*

In the March 1971 issue of this *Review*, James Melvin replied to my comment (in the same issue) on his 1968 treatment of the pattern of trade in a three-good, two-factor model. His reply, while conceding a basic error in his 1968 paper, takes issue with my analysis at several points. I reject his arguments on these points and herein attempt to show where his reasoning has gotten off the track.

In his 1968 paper Melvin makes the claim that whenever all goods are traded, the country exporting the labor intensive good will also be exporting the capital intensive good (p. 1263). In my comment I show that equal relative factor-endowments is a necessary and sufficient condition for his claim to be true. At the same time I point out that "the example from which Melvin generalizes is often a *possibility* when endowment ratios differ" (p. 241). In his reply Melvin maintains that in the quote above I provide a counterexample to the proposition I attempt to prove. He says, "Obviously if my claim is a possibility when endowment ratios differ, equal endowment ratios cannot be a necessary and sufficient condition for my claim" (p. 245). Melvin is in error here for I *have not* said his claim is a possibility when endowment ratios differ; I have said only that if endowment ratios differ a country *may or may not* export both the capital intensive and labor intensive goods. Melvin's claim implies it *must* export both. But, as I have shown, this is never the case. Thus it follows the necessary condition is equal endowment ratios.

The version of the Heckscher-Ohlin theorem put forth in my comment breaks the situation with production indeterminacy and unequal endowment ratios into two mutually exclusive cases: 1) each country must produce either the labor intensive or capital

intensive good, 2) one country's endowment ratio is the same as the optimal factor intensity of the good of intermediate intensity (denoted by  $k_3$  hereafter) and it produces only that good. Erroneously, Melvin asserts I have ignored a third case where one country exports both the labor intensive and capital intensive goods. Melvin might have a case were the pattern of trade the basis of categorization. Here, however, the basis is production, and if one country does not specialize in the good of intermediate intensity then each *must* produce one of the other goods; there is no other alternative. Thus, the two cases are exhaustive and include the case Melvin asserts is excluded.

It appears Melvin's misunderstanding above and his failure to see that my version of the Heckscher-Ohlin theorem goes substantially further than his own may in part be my own fault for not better explaining the use of the word "must" in my analysis. For production indeterminacy to exist it is necessary that both countries' endowment ratios lie between the optimal factor-intensities of the labor intensive and capital intensive goods. If a country's endowment ratio is greater than  $k_3$ , then it *must* produce the capital intensive good; it may produce all three goods or it may not produce either the labor intensive good or the good of intermediate intensity, but it *must* produce the capital intensive good. Similarly, a country *must* produce the labor intensive good if its endowment ratio is less than  $k_3$ . Clearly if two countries' endowment ratios lie on opposite sides of  $k_3$  we can say they must produce *different* goods, whereas if their endowment ratios lie on the same side of  $k_3$  they must produce the *same* good. This "same good" will be either the capital intensive or labor intensive good; it cannot be the one of intermediate intensity as Melvin suggests (p. 245, fn. 1).

\* San Diego State College.

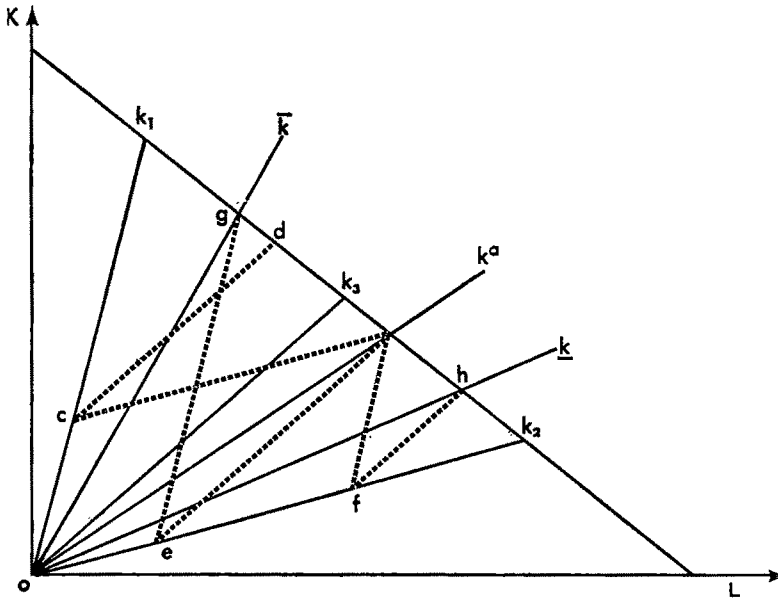


FIGURE 1

The wording of my version of the theorem should now be clear as it reads (in part): If both countries must produce the same good there is a minimum endowment-ratio difference above which that good must be exported by the country abundant in its intensively used factor, and if they must produce different goods there is a minimum endowment-ratio difference above which each country must export the good using intensively its abundant factor. Melvin's view of this is: "Unless precise conditions can be derived which tell us when it is *not* possible for one country to export both the capital intensive and labor intensive goods, we do not seem to have advanced much beyond my statement of the theorem" (his italics). Now these "precise conditions" must logically be the sufficient conditions for the standard Heckscher-Ohlin theorem.<sup>1</sup> But my version is easily stated in these terms: If endowment ratios lie on opposite sides of  $k_3$  a sufficient

condition for the standard theorem exists, otherwise<sup>2</sup> a sufficient condition does not exist.<sup>3</sup> Moreover, since the sufficient condition is in terms of a minimum endowment-ratio difference, it can be determined graphically as I show in the Appendix. Thus, contrary to Melvin's opinion, my version of the Heckscher-Ohlin theorem is very much along the line he, himself, suggests it should take.

In concluding his reply, Melvin questions my discussion of my Figure 3 (p. 242). He quite correctly points out the origin of the diagram need not be the origin for both countries. Thus, it is misleading of me to say "the country producing at  $R$  produces goods 1 and 2 in quantities represented by points  $b$  and  $d$ , respectively" for this could be interpreted to imply the distances from the origin to  $b$  and to  $d$  measure outputs of goods 1 and 2, respectively. This is incorrect. The point I was trying to make is that since point  $a$  is to

<sup>2</sup> This includes the situation in which one country's endowment ratio is equal to  $k_3$ .

<sup>3</sup> Figure 4 of my comment illustrates the case where both countries must produce the labor intensive good. Since  $oe$  and  $oj$  overlap there is no condition, short of arbitrarily restricting the production indeterminacy, which ensures the capital intensive good will be exported by the capital abundant country.

<sup>1</sup> Melvin implicitly goes beyond his own version of the Heckscher-Ohlin theorem when he says it is not surprising that as endowment differences increase a country becomes more likely to export the good using intensively its abundant factor. This result is not implied by his version.

the left of  $b$  and  $d$  is below  $c$ , the country producing at  $S$  must export good 2 and import good 1. Further, by the "similar tastes" assumption,<sup>4</sup> that same country must be the larger producer of good 2 and the smaller producer of good 1.

#### APPENDIX

Suppose Figure 1 represents a two-country trade equilibrium with production indeterminacy, and the ray  $k^a$  represents country  $A$ 's endowment ratio (notation as in my comment). Then  $oc$  and  $ef$  represent country  $A$ 's ranges of possible outputs of goods 1 and 2, respectively (in terms of production scale). Construct  $cd$  parallel to  $k_3$ . If country  $B$ 's endowment ray,  $k^b$ , passed through  $d$ ,  $oc$  would represent country  $B$ 's smallest possible output of good 1. Thus if  $k^b$  lies above  $od$ , country  $B$  must export good 1.

Construct  $eg$  parallel to  $k_1$ . If  $k^b$  passed

through  $g$ ,  $oe$  would represent country  $B$ 's largest possible production of good 2. Thus if  $k^b$  lies above  $og$ , country  $A$  must export good 2.

Let  $\bar{k}$  coincide with whichever of  $od$  and  $og$  is the steeper. Then for any  $k^b > \bar{k}$  the standard Heckscher-Ohlin theorem must hold. Similarly, let  $\underline{k}$  coincide with  $oh$  where  $fh$  is parallel to  $k_3$ . Then if  $k^b < \underline{k}$  the standard theorem may or may not hold, but country  $B$  must export good 2.

#### REFERENCES

- J. R. Melvin, "Production and Trade with Two Factors and Three Goods," *Amer. Econ. Rev.*, Dec. 1968, 58, 1249-68.
- , "Production Indeterminacy with Three Goods and Two Factors: Reply," *Amer. Econ. Rev.*, Mar. 1971, 61, 245-46.
- D. B. Stewart, "Production Indeterminacy with Three Goods and Two Factors: A Comment on the Pattern of Trade," *Amer. Econ. Rev.*, Mar. 1971, 61, 241-44.

<sup>4</sup> See fn. 2 of my comment.

# Production Indeterminacy with Three Goods and Two Factors: The Last Word?

By JAMES R. MELVIN\*

I am grateful for the opportunity of making this brief reply to Douglas Stewart's rejoinder. Let me first of all say that there has been some confusion between Stewart and myself about what he meant in his comment. I interpreted his statement that "equal relative factor-endowments is a necessary and sufficient condition for his claim to be true" (Stewart, (1972), p. 720), to mean: Equal relative factor-endowments is a necessary and sufficient condition for the possibility of one country exporting both the labor and the capital intensive goods. This apparently is not what Stewart meant, and I apologize for misinterpreting him.

I also seem to have interpreted Stewart as making a statement about trade when in fact he was concerned only with production. My misinterpretation can perhaps be explained, in part, by the fact that in a situation where we would expect to observe both countries producing all three goods,<sup>1</sup> I do not attach as much importance as Stewart obviously does to the conditions under which one country *must* produce one or the other of the goods.

In an attempt to avoid any further misinterpretations, I will resist the temptation to make further comments and will conclude this reply by simply expressing my views on the two points that I interpret as being at issue.

1) In a world of three goods and two factors it is possible, but by no means necessary, that one country will export both the capital intensive and the labor intensive goods.

2) In a two-good, two-factor case, and

given the usual assumptions, the Heckscher-Ohlin theorem is very powerful in that it is able to predict the pattern of trade with a knowledge of relative endowments only. For the three-good case, two theorems have been suggested.

*Melvin's version:* A country will export a bundle of commodities which uses the abundant factor most intensively.

*Stewart's version:* Excluding the possibility "that one country's endowment ratio is equal to the capital-labor ratio of the good of intermediate intensity and it produces only that good," (Stewart (1971) p. 243) and with  $k^a \neq k^b$ , "as  $|k^a - k^b|$  increases, the probability of each country exporting the good using intensively that country's abundant factor increases; if both countries must produce *different* goods, there is a value of  $|k^a - k^b|$  above which each country must export the good using intensively its abundant factor; if both countries must produce the *same* good, there is a value of  $|k^a - k^b|$  above which that good must be exported by the country abundant in its intensively used factor" (Stewart (1971) pp. 243-44).

I leave the decision as to which is the most informative version of the theorem to the reader.

## REFERENCES

- J. R. Melvin, "Production Indeterminacy with Three Goods and Two Factors: Reply," *Amer. Econ. Rev.*, Mar. 1971, 61, 245-46.
- D. B. Stewart, "Production Indeterminacy with Three Goods and Two Factors: A Comment on the Pattern of Trade," *Amer. Econ. Rev.*, Mar. 1971, 61, 241-44.
- , "Production Indeterminacy with Three Goods and Two Factors: Rejoinder," *Amer. Econ. Rev.*, Sept. 1972, 62, 720-22.

\* The University of Western Ontario.

<sup>1</sup> For any given equilibrium price line, of the infinity of possible production points, there are at most four which imply any kind of specialization.

# An Analysis of Turning Point Forecasts

By H. O. STEKLER\*

There have been a number of empirical analyses of the accuracy of economic forecasts (see Henri Theil, Victor Zarnowitz, and Stekler (1968a, b)). These studies have shown that economists have generally underestimated increases in *GNP* and its components. On the other hand, when these variables declined, the predictions frequently failed to forecast these movements, thus producing turning point errors. One of the studies (Stekler (1968a)) also indicated that forecasters were unable to "identify" cyclical peaks for some time after the occurrence of the event. For instance, while the peak preceding the 1960 recession occurred in the third quarter of 1960, none of the three sets of forecasts analyzed in this study predicted this turn in advance. The best performance was turned in by the forecasters who recognized the turn near the *end* of the third quarter. However, the forecasts of this set had, earlier in the third quarter, indicated that the economy would grow more rapidly in that quarter than it had in the second period. The other two sets of forecasts did not recognize the turn until the beginning or end of the fourth quarter of 1960.

Since we usually have no explicit information about the way economists generate forecasts, it would be desirable to attempt to find a mechanism which could reproduce many of the observed results.<sup>1</sup> This paper

\* Professor of economics, State University of New York at Stony Brook. I wish to thank Edward Ames, Susan Burch, and a referee for their valuable comments.

<sup>1</sup> This analysis is applicable to both judgmental and econometric forecasts because judgment is usually used to adjust the latter. Using a priori and judgmental information to adjust the econometric forecasts would tend to weaken the argument that econometric forecasts are scientific. The model builders, however, argue that the use of the forecasts obtained from the model without any personal judgment is a purely mechanistic forecasting procedure. The truth of the matter is that without these a priori adjustments, the forecasts of most models are not very accurate. In fact, the *ex ante* forecasts adjusted by the econometricians are generally much more accurate than the comparable *ex post* pre-

will focus on the turning point errors in the vicinity of cyclical peaks and troughs and will advance one hypothesis to explain why the observed errors might have occurred. This knowledge might prove useful in improving the quality of future forecasts.

## I. Predictions of Turning Points— A Hypothesis

In analyzing the prediction procedures that economists use in the vicinity of turning points, we shall assume that the forecasters begin with some subjective probabilities about the likelihood of a turn. As new information becomes available these subjective probabilities are then revised. Using this Bayesian type analysis, we shall attempt to determine whether forecasters fail to predict and recognize cyclical peaks because they do not "expect" such turns to occur; or stated another way, they have assigned low subjective probabilities to the possibility of a decline.

Let us assume that a forecaster assigns a subjective (prior) probability  $P(T)$ , to the likelihood of a cyclical turn; the prior probability that no turn will occur is, of course,  $P(NT) = 1 - P(T)$ . As the economist receives new information about the state of the economy, he would revise these probabilities. The probabilities which take into account the new evidence obtained about the state of the economy are called revised or posterior probabilities, and are denoted as  $P(T|S)$ , the probability of a turn given that a signal  $S$  from an indicator has been received;  $P(NT|S)$ , is the probability of no turn given a signal. Similarly  $P(T|NS)$  and  $P(NT|NS)$  refer to the revised probabilities if there is no signal from the indicator. It is possible to derive these revised probabilities using Bayes' Theorem.

For further discussions of these points see the paper by Michael Evans, Yoel Haitovsky, and George Treyz and the subsequent comment by this author.

$$(1) \quad P(T|S) \\ = \frac{P(S|T) \cdot P(T)}{P(S|T) \cdot P(T) + P(S|NT) \cdot P(NT)}$$

For purposes of analysis let us assume that a hypothetical forecaster examines one of the statistical indicators in preparing his forecasts. The Federal Reserve Board (*FRB*) Index of Production is a highly reliable indicator. Its movements are considered coincident with movements in the economy and is frequently used in analyzing the state of the economy. We shall therefore postulate that the forecaster examines this series for information about the state of the economy, and if this series reverses its previous movement, it is said to give a signal.<sup>2</sup> Some of these signals will be true and will coincide or lead cyclical turns and some will be false leads. The probability that a signal will occur when there is a turn is  $P(S|T)$ . Similarly the probability of a signal when no turn occurs is  $P(S|NT)$ .

Naturally, forecasters obtain information from more than one series prior to preparing a forecast, and the data available from these other series might conflict with the information obtained from the *FRB* Index. The analysis which is presented here would be considerably more complicated if the movements of more than one series and the associated joint probabilities were considered. There is an extremely large number of series which might be examined and there would be no theoretical justification for selecting some of these and not including the remainder in the analysis. In addition, if it can be shown that one series alone provides insight about the probability of a recession, that in itself is useful knowledge and should help explain why forecasters failed either to

predict or to identify cyclical peaks. Recognizing this caveat, the analysis will, therefore, assume that the forecaster acted as if he had only examined the *FRB* Index of Production.

## II. Results: 1957 and 1960 Cyclical Peaks

Using a retrospective analysis, it is possible to calculate the probabilities of a recession that a forecaster would have obtained near the 1957 and 1960 cyclical peaks. These probabilities can be calculated when movements in the *FRB* Index of Production are compared with the dates that the National Bureau of Economic Research has assigned to the business cycle turns. Suppose that in 1957 a forecaster had observed the postwar movements in the *FRB* Index relative to these business cycles. In the period 1947-56 there were 12 movements when the Index was below its previous peak for one or more months but which were not followed by recessions.<sup>3</sup> These movements occurred in 31 of the 95 months in which there was no recession.<sup>4</sup> Turns in the Index preceded or were coincident with both the 1948 and 1953 cyclical peaks. Consequently, in 1957 a fore-

<sup>3</sup> Historical data available in 1965 were used to construct these probabilities. Seasonally adjusted data were utilized and the data were obtained from *Business Statistics*, 1965 Biennial Edition. Although revisions in the index would have caused the levels of the *FRB* Index to differ from those which a forecaster might have observed in 1957, it is not likely that the number and length of the monthly movements would have been substantially altered. In practice, however, the preliminary data might show greater variability, but the forecaster must assume that these preliminary data reflect what the final figures will eventually reveal. The forecaster therefore would calculate his probabilities from the historical data, while also recognizing that a comparison of early and final data might show a number of turning point errors in the early data. (Although there have been comparisons of the early and final figures for a number of monthly series, I believe that no analysis of the *FRB* Index data has as yet been made.)

<sup>4</sup> This was a purely mechanical calculation and did not take into account major industrial strikes, which would cause declines in the index but might not produce a recession. To the extent that these false leads were eliminated, the calculated  $P(S|NT)$  would be decreased. However, in practice there is no a priori way of knowing which decline in the index would lead to a recession, and since we are analyzing an a priori approach, it would not have been appropriate to eliminate those movements.

<sup>2</sup> It is recognized that the *FRB* Index is only one of many series that an analyst examines. Furthermore, the importance given to the indicator, in this analysis, may be overstated since manufacturing's importance in the economy has been declining. On the other hand, the *FRB* Index, although still classified as a roughly coincident indicator, has tended to lead the rest of the economy. Consequently, if this series signals a decline, this prediction would be useful to the analyst in forecasting as well as identifying the subsequent cyclical movement.

caster might have constructed the following probabilities:

$$P(S|T) = 1.0 \quad P(S|NT) \cong .33$$

$$P(NS|T) = 0 \quad P(NS|NT) \cong .67$$

These probabilities are presented in Table 1. The prior probabilities that the economist has about the state of the economy must be taken into account, and Table 1 must, there-

TABLE 1—ESTIMATES OF RELATIVE FREQUENCY OF SIGNALS AND NO SIGNALS FROM THE *FRB* INDEX FOR TURNS AND NO TURNS IN ECONOMY, 1947–56

State of Economy	Information from <i>FRB</i> Index	
	Signal	No Signal
Turn	1.0	0
No Turn	.33	.67

fore, be modified. The new probabilities which are the product of the estimates contained in Table 1 and the economist's prior probabilities  $p$  and  $1-p$  are presented in Table 2. These are the relevant probabilities before a sample or new observation is obtained from the *FRB* Index of Production.

TABLE 2—PROBABILITIES OF SIGNALS AND NO SIGNALS IN *FRB* INDEX FOR TURNS AND NO TURNS IN ECONOMY, GIVEN PRIOR PROBABILITIES OF TURN AND NO TURN, 1947–56

State of Economy	Information from <i>FRB</i> Index	
	Signal	No Signal
Turn	1.00 ( $p$ )	0 ( $p$ )
No Turn	.33 ( $1-p$ )	.67 ( $1-p$ )

Suppose that in 1957 the economist's a priori estimate of the likelihood of a recession had been  $p = .20$ . This is the relevant subjective probability before obtaining any information from the *FRB* Index. These probabilities would be revised according to Bayes' Theorem (1) when the economist obtained information that the *FRB* Index had declined. According to (1), the revised probability  $P(T|S)$  after a decline would be .43. Since this posterior probability becomes

the prior probability relevant for the next sample, this procedure may be repeated. Using (1) the revised probabilities of a recession after observing the second and third months of decline in the *FRB* Index would have been .70 and .88, respectively. An examination of the data available on a contemporaneous basis in 1957 indicates that the seasonally adjusted *FRB* Index of Production had been below its December 1956 peak throughout 1957.<sup>5</sup> Consequently, Table 3 presents the revised probabilities of a recession that a forecaster might have had in early 1957, after allowing for the delay in publication of the data, if he had used this Bayesian approach to forecasting.

TABLE 3—REVISED PROBABILITIES OF A RECESSION IN 1957

Prior	Feb. 1957 <sup>a</sup> First decline observed	March 1957 Second decline observed	April 1957 Third decline observed
.10	.25	.50	.74
.20	.43	.70	.88

<sup>a</sup> This allows for the delay in publication. When the figures referring to January 1957 became available, the peak in December 1956 was identified. The January data became available in mid-February 1957, and were published in the February 1957 issue of the *Survey of Current Business*.

These probabilities are relatively high, even with a prior as low as  $p = .10$ . Since the downward movement in the Index persisted and the Index led the cycle, the forecaster should have had no difficulty in at least recognizing the 1957 recession when it began in July 1957, if he had used this approach and if his subjective probability had been  $p > 0$ .

The same procedure could have been applied in 1960 using data relating to the period 1947–59. The revised probabilities for the months leading up to the cyclical peak

<sup>5</sup> However, this does not assert that the index declined consecutively in each month of 1957, but merely that it was below its previous peak for each of those months. This was the criterion from which the probabilities of a downturn were calculated above. The revised data now indicate that the peak occurred in February 1957 rather than December 1956.

of May 1960<sup>6</sup> are presented in Table 4 and are virtually identical to those relevant for 1957. Consequently, similar conclusions hold for this period.

TABLE 4—REVISED PROBABILITIES OF  
A RECESSION IN 1960

Prior	March 1960 <sup>a</sup> First decline observed	April 1960 Second decline observed	May 1960 Third decline observed
.10	.26	.53	.78
.20	.44	.72	.89

<sup>a</sup> See Table 3, fn. a.

The failure to at least recognize the recessions of 1957 and/or 1960 might be attributable to at least 3 factors.

(i) The forecasters might not have examined the *FRB* Index in projecting *GNP*.

(ii) The forecasters might not have revised their predictions of *GNP* as the new *FRB* data became available.<sup>7</sup>

(iii) Their prior probabilities of a cyclical turn were zero.

Although forecasters may not have calculated the probabilities in exactly the manner that has been presented here, almost all forecasters look at both the current and recent values of the *FRB* Index and the historical relationship between movements in the Index and aggregate economic activity. Similarly, almost all forecasters revise their predictions as new information becomes available. While it is possible that other data might have yielded contradictory results, and consequently forecasters did not revise their forecasts for that reason, in the next section it will be shown that this was not likely in 1957 and 1960. Consequently, the first two factors are unlikely to have occurred. Factor (iii), that forecasters did not expect a recession and were "surprised" by its occurrence, is thus a plausible conclusion.

<sup>6</sup> This peak could only have been identified in July 1960, when the first data for June 1960 (which was below May) became available.

<sup>7</sup> They might have revised their forecasts but they might not have used the Bayesian approach that has been presented.

### III. Other Data for 1957 and 1960 and the Hypothesis

It was noted above that economists look at a variety of statistical data in constructing a forecast. In order to lend support to our hypothesis that the prior probability of a recession was very close to zero, we will show that a representative set of other data available contemporaneously with the *FRB* Index was not inconsistent with the movements of the *FRB* Index. For both 1957 and 1960 the contemporaneous movements of five other series were examined.<sup>8</sup> Two are considered leading series, average hours worked and new orders for durable goods; the other three, nonagriculture employment, personal income, and retail sales, are considered coincident series. Tables 5 and 6 show that the two leading series had their peaks around the same time as did the *FRB* Index and that throughout both periods they consistently remained below these peak levels.<sup>9</sup> Of the coincident series, only the personal income series moved to successive highs in the two periods, while the two other series displayed mixed patterns of strength and weakness. While the latter two series did not decline uniformly, even the two coincident series displayed some evidence of slowdowns, or weakness, in the nonagriculture employment data in 1957 and in both the nonagriculture employment and retail sale figures in 1960. Given these data, especially the leading indicators, and the movements of the *FRB* Index, if forecasters had "expected" the two recessions, even if they were not able to forecast them, they should at least have been able to identify them quickly.

### IV. Extensions to Other Turning Point Forecasts

If economists do not forecast downturns, a relevant question becomes: why are up-

<sup>8</sup> There was no data mining, i.e., the series were selected for a priori reasons, the data were collected, and all the results are reported. The data were obtained from the relevant 1957 and 1960 issues of the *Survey of Current Business*, i.e., the Date Reported heading of Tables 5 and 6 refer to the issue of the *Survey* from which the data were obtained.

<sup>9</sup> This analysis is based on the monthly change relative to a peak and not on the month-to-month movements, which tend to be erratic.

TABLE 5—PERCENT DECLINE FROM PEAK MONTH, AS REPORTED CONTEMPORANEOUSLY,<sup>a</sup>  
5 SELECTED SERIES, APRIL-SEPTEMBER 1957

	Average Hours Worked	New Orders— Durable Goods	Nonagriculture Employment	Personal Income	Retail Sales
Date Reported:					
April	—2.4 percent Dec. 1955	—10.1 percent Nov. 1956	P	P	—1.1 percent Feb. 1957
May	—2.7 percent Dec. 1956	—12.7 percent Nov. 1956	—0.1 percent Mar. 1957	P	—1.5 percent Feb. 1957
June	—3.2 percent Dec. 1956	—9.2 percent Nov. 1956	—0.0 percent Feb. 1957	P	P
July	—2.7 percent Dec. 1956	—16.8 percent Nov. 1956	—0.05 percent May 1957	P	P
August	—2.7 percent Dec. 1956	—17.2 percent Nov. 1956	P	P	P
September	—2.7 percent Dec. 1956	—17.2 percent Nov. 1956	—0.1 percent June 1957	P	P

<sup>a</sup> The month refers to the issue date of the *Survey of Current Business*, which usually publishes data for the immediately preceding month.

P—Peak in that particular month.

TABLE 6—PERCENT DECLINE FROM PEAK MONTH, AS REPORTED CONTEMPORANEOUSLY,<sup>a</sup>  
5 SELECTED SERIES, APRIL-SEPTEMBER 1960

	Average Hours Worked	New Orders— Durable Goods	Nonagriculture Employment	Personal Income	Retail Sales
Date Reported:					
April	—2.5 percent Dec. 1959	—2.2 percent Feb. 1960	—0.4 percent Feb. 1960	P	—0.3 percent Feb. 1960
May	—3.0 percent Dec. 1959	—2.7 percent Feb. 1960	P	P	P
June	—2.0 percent Dec. 1959	—1.2 percent Feb. 1960	—0.1 percent April 1960	P	—2.3 percent April 1960
July	—1.5 percent Dec. 1959	—4.1 percent Feb. 1960	—0.2 percent April 1960	P	—2.2 percent April 1960
August	—2.0 percent Dec. 1959	—5.0 percent Feb. 1960	P	P	—3.2 percent April 1960
September	—2.2 percent Dec. 1959	—3.4 percent Feb. 1960	—0.2 percent July 1960	P	—4.0 percent April 1960

<sup>a</sup> See notes to Table 5.

turns predicted? One explanation is that the upward momentum of a growing economy is recognized. A second possible explanation is

that the economists are aware that fiscal and monetary policy have been undertaken to combat the recession and they expect these

policies to bear fruition.<sup>10</sup> As evidence accumulates that these policies are succeeding, economists may or may not forecast the upturn, but at least they generally recognize the end of a recession.

In addition, this hypothesis might be used to explain the errors of the price forecasts for 1968-70.<sup>11</sup> In this period public policy was undertaken to slow down inflation. Since economists expected (i.e., the prior probability was high) these policies to succeed in slowing down price increases, every piece of data which might provide any evidence about the success of these policies was interpreted favorably.<sup>12</sup> Later the availability of more complete or revised data revealed the inaccuracy of these favorable interpretations and indicated that more time had to elapse before the policies were successful. In other words, expecting a set of policies to be successful could in this case have led to turning point errors of the other kind, i.e., predicting a turn when none occurred.

### V. Conclusions

On the basis of this evidence, we may conclude that the failure to predict or identify cyclical peaks may be attributable to the forecasters' extremely low subjective probabilities of the likelihood of such an event. Given this finding, what is the useful advice that can now be imparted to practicing forecasters? One suggestion is that they pay more attention to the statistical indicators, especially the leading series.

<sup>10</sup> Obviously, no one expects monetary and fiscal policy to cause a recession. Equally, this is naive faith in the wisdom of public policy.

<sup>11</sup> There has as yet been no general analysis of the 1968-70 GNP forecast errors. It would be interesting to determine whether this peak was again not predicted or "identified."

<sup>12</sup> A study of the factors, theories, and data which economists utilized in making their 1968-70 price predictions has just been begun.

At this stage in the development of forecasting techniques, it seems strange to advocate that forecasters examine the statistical indicators, especially the leading series, for these series have been utilized for a long time. However, as forecasters have come to emphasize quantitative forecasting techniques, less attention has been paid to the indicator approach. All the quantitative techniques however, require some judgmental elements, and it is just in this area that the leading indicators are most useful. They may provide some qualitative insights into the state of the economy which could influence a forecaster's judgment of the future and his subjective probabilities of the likelihood of a recession. This factor, in turn, might decrease the number of turning point errors and, thereby, improve the accuracy of economic forecasts.

### REFERENCES

- M. Evans, Y. Haitovsky, and G. I. Treyz, in "An Analysis of the Forecasting Properties of U.S. Econometric Models," in B. Hickman, ed., *Econometric Models of Cyclical Behavior*, Nat. Bur. Econ. Res. Stud. in Income and Wealth, vol. 36, New York 1972.
- H. O. Stekler, (1968a) "An Evaluation of Quarterly Judgmental Economic Forecasts," *J. Bus., Univ. Chicago*, July 1968, 41, 329-39.
- , (1968b) "Forecasting with Econometric Models: An Evaluation," *Econometrica*, July-Oct. 1968, 36, 437-63.
- H. Theil, *Applied Economic Forecasting*, Chicago 1966.
- V. Zarnowitz, *An Appraisal of Short-Term Economic Forecasts*, Occas. Paper 104, Nat. Bur. Econ. Res., New York 1967.
- U.S. Office of Business Economics, *Survey of Current Business*, various issues, 1957-60.
- , *Business Statistics*, 1965 Biennial ed., Washington 1966.

# The Permanent Income Hypothesis: Evidence From Time-Series Data

By PREM S. LAUMAS AND KHAN A. MOHABBAT\*

Milton Friedman's Permanent Income Hypothesis (*PIH*) appears to be one of the best known relationships that have been postulated between income and consumption. *PIH* in essence assumes proportionality between permanent income and permanent consumption. The most crucial assumption of the hypothesis, then, is that the transitory elements of income and consumption are uncorrelated; and, therefore, the marginal propensity to consume out of transitory income is zero. In addition, the *PIH* also assumes that the transitory and permanent elements of income and consumption are uncorrelated. Should the basic assumptions of *PIH* be true, it would yield policy implications which are very different from consumption functions based on the Absolute Income Hypothesis. Thus, for example, in a recent debate on economic policy the assumed validity of *PIH* has led to some far reaching conclusions. Friedman and David Meiselman, for example, argue strongly for the stability of monetary velocity compared with the Keynesian multiplier by using a concept of autonomous expenditures which in turn assumes a certain concept of income (national income minus taxes). They consider such a concept of income theoretically justifiable on the ground that it is closer to the definition of income required by Friedman's *PIH* (1964). Evidence to contradict the Friedman-Meiselman view has been advanced both for the United States and Canada (see, for example, Albert Ando and Franco Modigliani, Michael DePrano and Thomas Mayer, Donald Hester, Laumas and Gurcharan Laumas (1970),

and Laumas and Richard Zerbe). Since the issues raised in such discussions are of paramount importance it is imperative that further tests of *PIH* be made.

The purpose of this paper is to test the validity of the *PIH* with the help of U.S. data for the recent years. Specifically, we shall try to study: (a) the magnitude of the marginal propensity to consume out of transitory income as compared with permanent income; (b) the relationship between measured consumption and permanent and transitory incomes; and (c) the degree of correlation between transitory and permanent incomes.

Several tests of the *PIH* made in the past have attempted to do the same. Many, however, suffer from three major defects.

*First*, in most cases the authors have used one or only a few sources of transitory income. Thus, for example, Ronald Bodkin (1959) uses only one source; namely, the National Service Life Insurance Dividend paid to the veterans of World War II in the United States. Mordechai Kreinin also uses one source, that is, the restitution payments received by the Israelis from West Germany. Lawrence Klein and Nissan Liviatan use a few, such as life insurance benefits, gambling winnings, cash gifts, cash legacies, postwar credits, and other lump sum transfers of money. Obviously, the results based on one or a few sources of windfall income are not adequate to justify acceptance or the rejection of the *PIH*.

*Second*, in many cases the windfall income which has been considered by these authors as equivalent to Friedman's transitory income is not in fact transitory income for at least two reasons: (a) Windfall is basically a stock concept whereas transitory income is a flow concept. Failure to make this essential distinction has misled some writers (for example Bodkin (1959)) into rejecting the

\* Professor of economics and associate professor of economics, respectively, Northern Illinois University. We are thankful to an anonymous referee and to James A. Gherity and Gurcharan S. Laumas for helpful comments on an earlier draft of this paper. We are also thankful to our graduate assistant John Goveia for help in computations involved in this study.

*PIH*, and others into suggesting that it was a three-year moving average of measured income (e.g., James Tobin); (b) according to Friedman, transitory income arises due to accidental or chance factors and is, therefore, unanticipated. In the various studies made so far the windfalls were expected and, therefore, cannot be properly classified as transitory income. Thus, for example, the restitution payments had been the subject of prior negotiations and anticipated, by and large, by the recipients. Gambling winnings can also be regarded as permanent income because the gamblers are likely to count on them as a source of supplementary income. Heirs also are known to count on legacies; these would then be a final transfer of an asset that was a contingent asset for many years (Bodkin (1959)). Only, perhaps, in Bodkin's study the National Service Life Insurance dividends could be treated, from the expectation point of view, as transitory income properly. These dividends were announced in November 1949 and payment began on January 16, 1950. Because of the short time lag between the announcement of the payment and the actual payment such dividends can perhaps be said to be unanticipated by the recipients and, therefore, a transitory source of income.<sup>1</sup>

*Third*, the size of the transitory income relative to measured income is of some importance. Thus, for example, it has been suggested that the marginal propensity to consume is high for small windfalls but declines as the windfall increases. But in judging whether the receipts are large or small one must compare its size with the total receipts during a given period. In Michael Landsberger's view, the high marginal propensity to consume out of windfall income in Bodkin's study and the very low marginal propensity to consume out of windfall income in Kreinin's case can be reconciled by the fact that the windfall received by the veterans of World War II was only about 7 percent of their regular income—while the restitution payments received by the Israelis were as

large as their regular income. Bodkin (1966), however disagreed with this conclusion.

In this paper, we shall attempt to remove all these defects. We shall take into account all possible sources of transitory income and transitory consumption for the United States for several years. This is not easily possible in the cross-sectional analysis. So we use time-series data. However, in the following section we shall explain our method of calculating permanent income, permanent consumption, transitory income and transitory consumption before we present our results.

### I. Methodology

Friedman defined permanent income of a consumer unit as the product of an interest rate  $r$  and a stock of wealth  $W$ ; the stock of wealth is to be interpreted as the present value of anticipated future receipts from both human and nonhuman assets discounted back to the present at an objective rate of interest whose average value is  $r$ .<sup>2</sup> Permanent income is thus a theoretical construct. For the aggregate data, however, estimates of permanent income can be made by considering it as the weighted average of the past values of measured income, the weights declining exponentially through time. Both the weights and the number of years are allowed to be determined by the data, the weights by the multiple correlation and the number of years by adding years successively until an additional year produces no significant increase in the coefficient of correlation.<sup>3</sup> A discrete analogue of the following equation was fitted

$$(1) \quad Y_p(T) = \beta \int_{-\infty}^T e^{(\beta-r)(t-T)} Y(t) dt$$

where  $Y_p$  is permanent income,  $Y$  is measured income,  $T$  is the date for which estimate is constructed, and  $t$  covers the whole range of earlier dates;  $\beta=r$ , the rate of interest. If the consumers have short horizons they will weigh receipts in the near future more heavily compared with the distant re-

<sup>1</sup> See, however, Friedman's disagreement on this point with Bodkin in his "Comment" on Bodkin's paper (1960).

<sup>2</sup> Compare Friedman (1963).

<sup>3</sup> For further details of this method see Friedman (1957), pp. 142-47.

ceipts and vice versa if they have long horizons. In the former case they use a higher interest rate compared with the latter. Then,  $1/\beta$  can be regarded as the horizon or the "number of year's purchase" corresponding to the rate of interest. The trend factor is  $\alpha$ . Without the trend factor, equation (1) when applied to a steadily growing series yields estimated values of permanent income systematically below their measured values. In other words it does not take into account any past rate of increase in wealth. In order to remove this defect we adjust the relevant permanent income and permanent consumption series by a trend factor,  $\alpha$ . Following Friedman, the value of  $\alpha$  was taken to be .02 for both the income and consumption series.

By taking several different values of  $\beta$  alternative  $Y_p$  series were made to compute the consumption function. By successively substituting one such series for another in the consumption function we finally chose disposable income with  $\beta=.4$  for the consumption series including durable goods. The  $\beta$  weight also turned out to be .4 for the consumption series excluding durable goods. With  $\beta=.4$ , sixteen years (1929-48 excluding 1942-45 war years)<sup>4</sup> were lost in computing permanent income. With  $\beta=.4$ , the weights used were .330, .221, .148, .099, .067, .045, .030, .020, .013, .009, .006, .004, .003, .002, .001, .001, .001. The sum of these weights equals 1. It is interesting to note that Friedman's  $\beta$  weight for a much longer period (1905-51) for the United States was exactly the same. This indicates a relative constancy of the horizon in spite of several structural changes in the economy over time.

Friedman does not offer any specific method of computing permanent consumption. He only suggests that permanent consumption is highly correlated with permanent income. Consequently, several series of permanent consumption with durable goods and without durable goods, with different

values of  $\beta$ , were computed. Finally, we related each one of these series with the permanent income series which showed the highest coefficient of correlation. Permanent consumption series with  $\beta=.4$ , with durable goods and  $\beta=.5$  without durable goods turned out to be the relevant series. In order to take the maximum period into account consumption functions were fitted for the period 1949-70, corresponding to  $\beta=.4$  for disposable income mentioned above.<sup>5</sup>

Transitory components of income and consumption were found by deducting the permanent components from their respective measured values.

## II. Results

Consumption, as defined in the *PIH* (consumption of nondurable goods plus the rental value of durable goods) is very hard to compute for the United States due to the lack of relevant data. Consequently, two linear regressions were computed, one with consumption including the expenditures for durable goods and the other excluding them. The first gives too high a measure of consumption from the viewpoint of the *PIH* and the second too low. We assume that the values intermediate between them will reflect the true values from the standpoint of the *PIH*. Table 1 presents the results of this study.

The information in Table 1 indicates that if due allowance is made for the use of durable goods,  $MCP_p$  would lie between .779 and .906 (closer to .779) and the  $MPC_t$  between .554 and .835. Assuming that the true  $MPC_p=.82$  and the true  $MPC_t=.64$ , one can conclude that Friedman's *PIH* in its strict form does not hold. The  $MPC$  out of measured income (including durable goods) was approximately equal to .901. Thus the marginal propensity to consume out of transitory income is roughly 2/3 of the  $MPC$  out of measured income.

In addition to the above test a related direct test of the *PIH* was made in which mea-

<sup>4</sup> War years were excluded on the ground that special circumstances of this period made it absurd to estimate an equation like (1) to calculate permanent income. Moreover, during this period the consumption data had abnormal transitory elements, compare Friedman (1957) p. 146.

<sup>5</sup> It may be noted that this method of calculating the permanent value of a variable is now in common use. For some of its interesting applications, see Friedman (1959), George Morrison, and Laumas and Laumas (1969).

TABLE 1—MARGINAL PROPENSITIES TO CONSUME  
PERMANENT AND TRANSITORY INCOME IN THE  
UNITED STATES (1949–1970)

					Computed <i>t</i> -Value of the Regression Equation
	<i>R</i>	<i>D.W.</i>			
$C_p^* = 3.443 + 0.906Y_p$ [8.902] (.003)	.99	.448 <sup>c</sup>			296.475
$C_p = 7.551 + 0.779Y_p$ [17.181] (.003)	.99	.303 <sup>c</sup>			224.420
$C_t^* = 1.915 + 0.835Y_t$ [4.979] (.026)	.99	1.849 <sup>a</sup>			31.702
$C_t = -0.216 + 0.554Y_t$ [-0.740] (.019)	.98	1.067 <sup>b</sup>			27.816

Source: *Survey of Current Business*, various issues.

Note: Figures in parentheses indicate standard errors of the regression coefficient. Figures in brackets are the *t*-values for the intercepts. Intercepts of all regression equations except that of the marginal propensity to consume transitory income without durable goods are significant at the 5 percent level.

$C^*$ =aggregate consumption including durable goods.

$C$ =aggregate consumption excluding durable goods.

$R$ =Coefficient of correlation.

$D.W.$ =Durbin-Watson.

Subscripts *p* and *t* represent permanent and transitory variables.

<sup>a</sup> denotes no serial correlation at 5 percent level of significance.

<sup>b</sup> denotes no serial correlation at 1 percent level of significance.

<sup>c</sup> denotes serial correlation. (However, this only reduces the efficiency of Ordinary Least Square estimators. Fortunately, the *t*-values of the regression equations are quite large.)

sured consumption is the dependent variable and the transitory and permanent components of income the explanatory variables. Estimates were made for the following equation.

$$(2) \quad C^* = \beta_0 + \beta_1 Y_p + \beta_2 Y_t$$

The results are presented in Table 2 and support the estimates from Table 1. For equation (2) the Durbin-Watson test at 2.5 percent level shows no serial correlation. The intercept of this equation as well as the regression coefficients are significant at the 1 percent level.

TABLE 2—MARGINAL PROPENSITIES TO CONSUME  
PERMANENT AND TRANSITORY INCOMES FOR  
THE UNITED STATES (1949–1970)

	<i>R</i>	<i>D.W.</i>
$C^* = 3.766 + 0.926Y_p + 0.655Y_t$ [6.401] (0.010) (0.086)	.99	1.510

Note: For explanation of symbols, see Table 1.

In addition to the above calculations, we tested Friedman's assumption of zero correlation between transitory income and permanent income. The coefficient of correlation in this case was .885. This additional evidence appears in line with the previous results.

### III. Concluding Remarks

On the basis of the above results, we feel that for the recent years for the United States the findings of consumption functions based on Friedman's *PIH* cannot be accepted without reservations. Some other studies also indicate similar results though somewhat less damaging to the *PIH*.<sup>6</sup>

In conclusion, however, we must add that although we have followed Friedman very faithfully in his method for calculating the permanent value of a variable, we do not fully agree that his methodology corresponds most accurately to his theoretical construct of permanent income. True, there is a considerable amount of valid psychological reasoning about consumer behavior behind this method,<sup>7</sup> yet it appears to us that the assumption that the biggest response occurs immediately at the beginning of the adjustment period is rather strong. This would be a particularly unfortunate assumption, in some cases, when using monthly or even quarterly data.<sup>8</sup>

### REFERENCES

A. Ando and F. Modigliani, "Velocity and Investment Multiplier," *Amer. Econ. Rev.*, Sept. 1965, 55, 693–728.

<sup>6</sup> See, e.g., Laumas (1969), Bird and Bodkin, H. W. Watts, and Robert Eisner.

<sup>7</sup> Compare Simon (1966), and Laumas and Laumas (1971).

<sup>8</sup> Reasoning of this type has led Zvi Griliches to conclude that, "the theoretical rationalizations offered [for such models] are often skin deep."

- R. C. Bird and R. G. Bodkin, "The National Service Life Insurance Dividend of 1950 and Consumption: A Further test of the 'Strict' Permanent Income Hypothesis," *J. Polit. Econ.*, Oct. 1965, 73, 499-515.
- R. G. Bodkin, "Windfall Income and Consumption," *Amer. Econ. Rev.*, Sept. 1959, 49, 602-14.
- , "Windfall Income and Consumption: Reply," *Amer. Econ. Rev.*, June 1966, 56, 540-46.
- M. DePrano and T. Mayer, "Tests of the Relative Importance of Autonomous Expenditures and Money," *Amer. Econ. Rev.*, Sept. 1965, 55, 729-52.
- R. Eisner, "The Permanent Income Hypothesis: Comment," *Amer. Econ. Rev.*, Dec. 1958, 48, 972-90.
- M. Friedman, *A Theory of Consumption Function*, Princeton 1957.
- , "The Demand for Money: Some Theoretical and Empirical Results," *J. Polit. Econ.*, Aug. 1959, 67, 327-51.
- , "Comment on Bodkin's Paper," in I. Friend and R. Jones, eds., *Study of Consumer Expenditures, Incomes and Savings*; Proc. Conference on Consumption and Saving, Univ. Pennsylvania 1960.
- , "Windfall, the 'Horizon,' and Related Concepts in Permanent Income Hypothesis," in C. F. Christ et al. eds., *Measurement in Economics: Studies in Mathematical Economics and Econometrics*, Stanford 1963, 3-28.
- , "Reply to Donald Hester," *Rev. Econ. Statist.*, Nov. 1964, 46, 369-76.
- and D. Meiselman, "The Relative Stability of Monetary Velocity and Investment Multiplier in the United States, 1897-1958" in E. C. Brown et al. eds., *Stabilization Policies*, New York 1963, 165-268.
- Z. Griliches, "Distributed Lags: A Survey," *Econometrica*, Jan. 1967, 35, 16-49.
- D. Hester, "Keynes and the Quantity Theory: A Comment on the Friedman-Meiselman CMC paper," *Rev. Econ. Statist.*, Nov. 1964, 46, 364-68.
- L. R. Klein and N. Liviatan, "The Significance of Income Variability on Savings Behavior," *Bull. Oxford Univ. Inst. Econ. Statist.*, May 1957, 19, 151-60.
- M. Kreinin, "Windfall Income and Consumption," *Amer. Econ. Rev.*, June 1961, 51, 388-90.
- M. Landsberger, "Windfall Income and Consumption: Comment," *Amer. Econ. Rev.*, June 1966, 56, 534-40.
- P. S. Laumas, "A Test of the Permanent Income Hypothesis," *J. Polit. Econ.*, Sept./Oct., 1969, 77, 857-61.
- and G. S. Laumas, "Interest-Elasticity of Demand for Money," *Southern Econ. J.*, July 1969, 36, 90-93.
- and ———, "The Definition of Money and the Relative Importance of Autonomous Expenditures and Money," *Metroeconomica*, Apr. 1970, 22, 88-99.
- and ———, "On How to Calculate Permanent Income," mimeo. 1971.
- P. S. Laumas and R. O. Zerbe, "The Relative Stability of Monetary Velocity and Investment Multiplier in Canada," *J. Money, Credit, Banking*, Nov. 1971, 3, 867-77.
- G. Morrison, *Liquidity Preference of Commercial Banks*, Chicago 1966.
- H. A. Simon, "A Note on Jost's Law and Exponential Forgetting," *Psychometrika*, Dec. 1966, 31, 505-06.
- J. Tobin, "On a Theory of Consumption Function" in L. H. Clark, ed., *Consumer Behavior*, Vol. 2, New York 1958, 447-54.
- H. W. Watts, *An Analysis of the Effects of Transitory Income on Expenditure of Norwegian Households*, Cowles Foundation Disc. pap. 149, New Haven 1962.

# The Incidence of the Social Security Payroll Tax: Comment

By MARTIN S. FELDSTEIN\*

In a recent paper in this *Review*, John Brittain claims to have shown that "the real burden of the tax falls on labor" (p. 122). The purpose of this note is to show that his claim is unwarranted, that no implication about the incidence of the tax can be drawn from either his theoretical discussion or his empirical analysis.

## I

A full assessment of the long-run incidence of the social security tax requires answers to four questions. 1) How much does the tax alter the quantities of labor and capital supplied? 2) How do changes in factor supplies affect the marginal products of labor and capital? 3) How is the wage rate and the return on capital related to these marginal products? 4) How does the tax affect the relative prices of the goods consumed disproportionately by labor and by the owners of capital? Brittain's theoretical discussion deals with the supply of labor. His empirical analysis is concerned only with questions 2) and 3) and provides no information about the effect of the tax on factor supplies.

Most analyses of the incidence of the payroll tax concentrate on the tax's effect on the supply of labor. A more elastic aggregate labor supply generally implies that, *ceteris paribus*, a smaller fraction of the burden falls on labor. Brittain concludes from his theoretical discussion that the tax does not change the quantity of labor supplied, and that the burden of the tax therefore falls on labor. Because of the tax's effects on the capital stock and on relative prices, the second does not follow even if the first is true. However, I will concentrate on examining the basis of Brittain's conclusion that the quantity of labor is unaffected by the tax.

If the aggregate labor supply were com-

pletely inelastic with respect to the wage rate, it is obvious that it would also be unaffected by the tax. The important question is the effect of the tax if the labor supply is elastic. Brittain concludes that in this case the tax will not change the quantity of labor supplied if "labor bargains in terms of total compensation" (p. 114). The assumption that labor bargains in terms of total compensation implies that the quantity of labor supplied at each *gross* wage is unaffected by what fraction of the gross wage is paid in taxes, and that any tax change is therefore ignored by both employer and employee.

Brittain accepts this extremely implausible assumption as the basis for his strong conclusion because of a more general error in his analysis. Although the question at issue should be the incidence of the entire payroll tax, Brittain concentrates his attention on the share paid by the employer. Moreover, he implicitly assumes that the *employees'* share of the tax is viewed by them as equivalent to income and therefore entirely borne by labor! He then concludes that, since it would be irrational to treat the two parts of the tax differently, the *employers'* share must also be borne by labor.

More specifically, Brittain states that the employers' share of the tax would not be fully borne by labor only "if the supply curve of labor were not perfectly inelastic *and* the supply price excluded the employer's tax . . ." (p. 115). He then rejects the latter condition in favor of the view that labor is indifferent between a higher net wage and a higher employer tax contribution. He argues that to believe otherwise "depends on labor viewing one withheld tax as part of its income but not the other. This behavior (is) difficult to rationalize . . ." (p. 115). That is, since income taxes and the employees' portion of the payroll tax are, by implicit assumption, treated as net income and therefore borne by

\* Professor of economics, Harvard University.

labor, the employers' contribution must also be.<sup>1</sup>

Such an argument for assuming that labor does not treat the payroll tax as a reduction in its income is clearly unacceptable. More generally, the key question of the effect of the tax on labor supply cannot be resolved without empirical analysis.

Brittain's paper also ignores the effect of the social security program on the accumulation of capital. In fact, by reducing disposable income the payroll tax is likely to lower saving. Perhaps more important, the old age retirement benefits are likely to reduce the incentive to save. Such reductions in saving would lower the long-run capital stock, depressing wages and raising the rate of interest. Thus even if the labor supply is constant, the capital formation effect of the social security program is likely to depress real wages.

Because Brittain has implicitly assumed a one-good economy, he makes no mention of the distributional effects of changes in relative prices. A full analysis of the incidence of this tax should recognize the possibility of forward shifting and consider how income elasticities differ among more and less labor-intensive goods.

## II

Brittain claims that his empirical analysis shows that the entire payroll tax is borne by labor. In fact, his regressions provide no basis for such a conclusion. There is no evidence about the effect of the tax on the supplies of labor and capital nor about its affect on gross real factor prices.

Brittain's analysis actually deals with the second and third questions raised in the beginning of Section I above. More specifically, the estimates of the elasticity of substitution between capital and labor that could be derived from the production function regressions show how changes in the capital-labor

ratio would alter relative marginal products.<sup>2</sup> In addition, as the following discussion shows, the estimates of the parameter  $s$  in Brittain's regressions can be interpreted as evidence that the gross wage rate is equal (or at least proportional) to the marginal product of labor. Brittain is not justified in interpreting this coefficient as a measure of shifting.

The correct interpretation of Brittain's  $s$  parameter can be easily obtained. Brittain's empirical study is based on the assumption that the technology of an industry in different countries can be approximated by an aggregate constant elasticity of substitution (CES) production function:

$$(1) \quad V = (\alpha K^{-\rho} + \beta L^{-\rho})^{-1/\rho}$$

This implies that the marginal product of labor is

$$(2) \quad \frac{\partial V}{\partial L} = \beta \left[ \frac{V}{L} \right]^{1/\sigma}$$

where  $\sigma = (1 + \rho)^{-1}$  is the elasticity of substitution. If the neoclassical assumption that firms pay a gross wage equal to the marginal product of labor is correct, equation (2) implies

$$(3) \quad W(1 + t) = \beta(V/L)^{1/\sigma}$$

where  $W$  is the net wage to labor and  $t$  is the proportional rate of payroll tax.

Taking logarithms and rearranging yields:

$$(4) \quad \log W = \log \beta + \frac{1}{\sigma} \log (V/L) - \log (1 + t)$$

For relatively small tax rates,  $\log (1 + t)$  can be approximated by the linear term of the Taylor's expansion, i.e., by  $t$ ; Brittain uses this approximation. Equation (4) can then be written

$$(5) \quad \log W = \log \beta + \frac{1}{\sigma} \log (V/L) - t$$

<sup>1</sup> Although there is a brief footnote acknowledging that the employee share of the tax might also be excluded from the supply price of labor, Brittain does not appear to recognize the significance of disregarding this possibility in his general analysis.

<sup>2</sup> The regression coefficients from which to derive the estimated elasticities of substitution are not, however, actually presented in Brittain's article.

Note that this relation contains no information about the response of labor supply or capital stock to the tax rate. It cannot by itself provide any information about tax shifting. It takes the values of  $V/L$  in different countries as given when the essence of the tax shifting problem is to assess how the tax changes  $K$ ,  $L$  and the marginal product of labor.

Nevertheless, equation (5) is the basis of Brittain's empirical analysis. He introduces what he calls a "shifting parameter,"  $s$ , as a coefficient of  $t$ :

$$(6) \quad \log W = \log \beta + \frac{1}{\sigma} \log (V/L) - st$$

He finds that the estimates of  $s$ , which he interprets as the proportion of the tax that is shifted, are almost never significantly different from one. This interpretation is incorrect. The derivation of equation (6) shows that if the assumptions used to justify the regression are correct, the estimated value of  $s$  should be identically equal to one no matter how much of the tax is shifted.<sup>3</sup> That is, the value of  $s$  in equation (5) is implicitly one even though no assumptions about the extent of shifting have been made. A value of  $s$  different from one would be evidence against the assumptions of a CES technology with labor paid its marginal product. Unless the effect of the tax on  $V/L$  is known, the finding that  $s$  equals one provides no information about the extent of actual shifting.<sup>4</sup>

<sup>3</sup> In practice, of course, the estimated values of  $s$  were not identically equal to one. There are several reasons for this. 1) The assumption of an aggregate CES technology with the same parameters in different countries is an approximation. 2) The equality of the gross wage and the marginal product of labor would be disturbed by market imperfections. 3) A stochastic element in the production function would be a source not only of variation in the estimated parameters but, because of the simultaneity of the system, of bias as well. 4) Brittain used an approximation to represent  $\log (1+st)$  as a linear function of  $s$ . 5) The data on  $V$ ,  $L$ ,  $W$ , and  $t$  do not correspond exactly to the quantities of economic theory. Brittain's estimates of  $s$  and the overall goodness of fit of his regression equations are essentially an appraisal of the seriousness of these five econometric problems.

<sup>4</sup> The value of  $s$  estimates the proportion of the tax falling on labor only if technology is CES and the tax

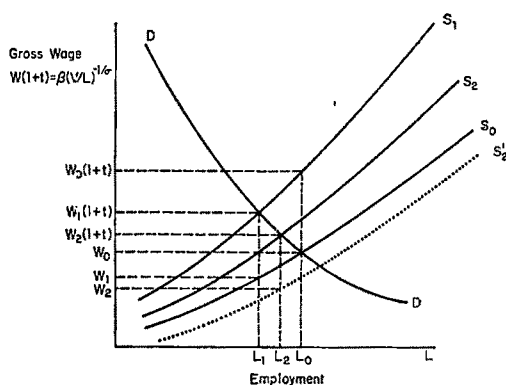


FIGURE 1

### III

The estimated value of  $s$  also provides no information about the extent to which the labor supply curve shifts in response to the tax. More specifically, the value  $s=1$  does not distinguish the case in which "labor bargains in terms of total compensation" (i.e., in which the supply depends only on the gross wage) from the case in which only the net wage matters. Because Brittain has estimated only the parameters of a labor demand curve, his estimates do not reflect either the elasticity or the shifting of the supply curve.

Figure 1 shows that a value of  $s$  equal to one is consistent with any shifting of the labor supply curve and, more generally, any incidence of the tax. The  $DD$  curve shows the demand for labor, i.e., the marginal product of labor as a function of the level of employment.<sup>5</sup> At every point on  $DD$ , the gross wage  $W(1+t)$  is equal to the corresponding marginal product of labor  $\beta (V/L)^{1/\sigma}$ . This  $DD$  curve is the relation that Brittain has estimated; his values of  $s=1$  and his high  $R^2$ s provide support for the assumptions of CES technology and neoclassical factor payments that underlie the  $DD$  curve. Supply curve  $S_0$  is the supply curve before the in-

does not alter  $V/L$ , i.e., the tax leaves the relative supply of labor and capital unchanged. For practical purposes, this requires that both labor and capital are unchanged by the tax.

<sup>5</sup> For simplicity, the capital stock is assumed constant.

roduction of the tax. The original equilibrium is therefore with employment  $L_0$  and wage  $W_0$ .

Now let a proportional tax at rate  $t$  be introduced. If employees care only about the net wage, their supply curve in terms of the gross wage shifts upwards by a factor of  $(1+t)$  to  $S_1$ . Production occurs with the gross wage indicated by the intersection of  $S_1$  and  $DD$ , i.e.,  $W_1(1+t)$ . Note that the net wage  $W_1$  is equal to the marginal product of labor divided by  $1+t$ .

Consider now the other case that Brittain discusses: the employees consider some of the tax to be income to themselves so the new supply curve  $S_2$  is actually shifted up by a factor of  $1+\theta t$  for  $0 \leq \theta < 1$ .<sup>6</sup> The new equilibrium is at a level of employment  $L_2$  that is between  $L_0$  and  $L_1$ . The marginal product at  $L_2$  is equal to the gross wage  $W_2(1+t)$ . (The net wage  $W_2$  corresponds to a point below the  $S_0$  line; supply curve  $S_2'$  relates the supply of labor to the net wage in the presence of the tax.) In terms of Brittain's analysis, the important point to recognize is that at this equilibrium point, as at every equilibrium, the marginal product of labor  $\beta(V/L)^{1/\sigma}$  is still equal to  $1+t$  times the corresponding net wage. This is what Brittain's regression confirms. It does not depend on how much the supply curve shifts. The value of  $s$  is equal to 1 regardless of the value of  $\theta$ .<sup>7</sup>

#### IV

The incidence of the social security tax is a question of substantial and growing importance. Brittain has shown that the inter-

country data is consistent with a model in which industry production functions have constant elasticity of substitution and in which the gross wage is equal to the marginal product of labor. Unfortunately, his paper provides no evidence on the response of labor supply or the incidence of the tax. His conclusion that the entire burden of the tax is borne by labor is not supported by either the theoretical discussion or the regression analysis. At most, his empirical analysis has shown that, *if the supply of labor and capital were not changed by the tax*, the working of the factor markets would cause the entire burden to fall on labor.

A proper assessment of the incidence requires an estimate of a simultaneous equations model of the supply and demand for both labor and capital. Ideally, the study should also show the effects of the tax on the relative prices of the goods consumed disproportionately by labor and the owners of capital. Brittain only estimates the demand equations. Even preliminary conclusions will be possible only after we have estimates of the changes in the supply of labor and capital in response to changes in the social security program.

The measure of labor supply in both the demand and supply studies should go beyond the number of workers to deal with occupational choice and individual labor effort.<sup>8</sup> Moreover, the fact that the social security program is financed by a tax on labor should not be allowed to obscure the possible importance of its effects on the supply of capital.

#### REFERENCE

- J. Brittain, "The Incidence of Social Security Payroll Taxes," *Amer. Econ. Rev.*, Mar. 1971, 61, 110-25.

<sup>6</sup> In the extreme case in which "labor bargains in terms of total compensation,"  $\theta=0$  and the new supply curve is the same as  $S_0$ . Note that while Brittain's evidence is consistent with this, it is also consistent with any other value of  $\theta$ .

<sup>7</sup> It might be objected that this discussion is too closely tied to the CES production model. Brittain suggested that the regressions of  $\log W$  on  $\log(V/L)$  and the tax variable were interesting outside the context of that model. But no other justification for such a regression is offered and it is not clear what possible interpretation could then be given to the estimated value of  $s$ . The logarithm of the marginal product of labor is related to the logarithm of  $V/L$  only if the technology is CES.

<sup>8</sup> General estimates of the elasticity of labor supply with respect to net wages are not sufficient for studying this question because of the link between individual benefits and tax contributions; i.e., the incidence of the program is affected by the shift in the supply curve as well as by its elasticity. The analysis must also reflect such special characteristics of the social security tax as the ceiling on taxable income and the full taxing of married women workers without a corresponding benefit increase.

# The Incidence of the Social Security Payroll Tax: Reply

By JOHN A. BRITTAIN\*

A casual reading of Martin Feldstein's stern critique might suggest that my analysis is about 200 percent wrong. On a more conciliatory note, I judge Feldstein to be less than 100 percent in error. However, my appreciation of his well-meaning efforts to state the assumptions of my analysis for me and to reinterpret the results is minimal; in short, I am not at all tempted to lie in the Procrustean bed he has so laboriously made up. Feldstein charges me with insufficient generality and a flaw in a priori reasoning; but the nub of his criticism involves his interpretation of the "shifting" coefficient  $s$  in my empirical formulation. He dismisses on two grounds my empirical finding that  $s$  is equal to unity; he asserts that 1) this result is built-in by the assumptions of neoclassical theory, and 2) it carries no implications for the "incidence" of the tax. The essence of the Feldstein maneuver is to derive by assumption (his assumptions, not mine) what I have sought to estimate empirically and then to assert that my analysis is a test of these assumptions, rather than an analysis of incidence. A discussion of these points in turn will raise also the related issues taken up by Feldstein.

## I. Interpretation of the Model and the Coefficient $s$

My point of departure was the linear estimating equation relating the logarithms of labor productivity and the real wage rate, as originally specified by *ACMS*.<sup>1</sup> As I stated, and Feldstein showed explicitly, one underlying rationale for this relationship is the assumption that labor is paid its marginal product in the context of the *CES* model. Although I asserted (p. 117) that the validity of the analysis is not dependent on these rigid assumptions, this is not the issue to be debated here.<sup>2</sup> Indeed, as I said at the be-

ginning (p. 118), the question which immediately arises with respect to the estimating equation is whether the wage rate associated with a given level of productivity includes the employer tax per unit of labor, in addition to the nominal wage  $w$ .<sup>3</sup> For investigation of this question, the rate of compensation per worker was specified generally as  $w$  plus some unknown fraction  $s$  of the employer payroll tax  $tw$ . The generalized version of the *ACMS* equation then reads:

$$(1) \quad \log w(1 + st) = a + b \log V/L$$

This is essentially Feldstein's relation (3) after the logarithmic transformation, *except* that Feldstein has taken the liberty of dropping the  $s$  from my formulation.<sup>4</sup>

Having chosen to respecify my equation for his own purpose, Feldstein asserts that "if the assumptions used to justify the re-

productivity was a priori plausible, and even the logarithmic transformation can be justified on the empirical ground of improved homoscadasticity. However, Feldstein says that it is not clear what interpretation can be given to the estimated coefficient  $s$  in the absence of the *CES* and marginal productivity assumptions. I suggested such an interpretation, given the assumption of minimal employment effects (p. 118, including fn. 25). Nevertheless, for the purpose of the present discussion I am willing to let Feldstein hold me to *CES* and the equation of some variant of wages to marginal productivity; this is not to imply, however, that the meaning of the findings is necessarily dependent on these assumptions.

<sup>3</sup> I stressed the employer tax because of the greater controversy surrounding its incidence. Not everyone accepts the theorem that the effects of a tax are independent of the side of the labor market on which it is imposed. The prevalent assumption in the social security literature appears to be that the employee tax is borne by labor, but that the incidence of the employer tax is in doubt. In any case, in stressing the employer tax, I assumed that if it were found that the part of the tax nominally paid by the employer actually comes out of labor income, no one would doubt that the employee tax is also borne by labor.

<sup>4</sup> Feldstein's formulation also differs in that his  $W$  is not the nominal wage rate, since it is net of both employer and employee taxes; his tax rate  $t$  is also the sum of the two rates. However, this difference is irrelevant to my complaint that Feldstein has assumed  $s=1$ .

\* Senior Fellow at the Brookings Institution.

<sup>1</sup> See Kenneth Arrow et al, p. 228.

<sup>2</sup> I suggested that a relationship between wages and

gression are correct, the estimated value of  $s$  should be identically equal to one . . ." (p. 737). This is simply incorrect. I indicated the CES and marginal analysis framework was one rationale for my regression, but I did not assume  $s=1$ ; Feldstein did that for me. An even more remarkable misinterpretation is Feldstein's assertion that *his* equation (5), which excludes my coefficient  $s$ , "is the basis of Brittain's empirical analysis" (p. 737). This is obviously untrue, since  $s$  was included explicitly everywhere as the key parameter to be estimated, and all of the empirical results reported are estimates of this coefficient which Feldstein deleted from my model. Feldstein also seems to be saying that I introduced  $s$  as an afterthought upon arriving at his equation (5). I repeat that the latter is his equation, not mine, and that I introduced  $s$  and the tax rate  $t$  together at the outset, according to the rationale for equation (1) above. I did not replace  $t$  by  $st$ , as Feldstein sees it; rather, in his equation (3), he replaced my  $st$  by  $t!$

Feldstein sets my coefficient  $s$  equal to unity on the assumption that firms pay a "gross wage" (inclusive of all payroll taxes) equal to the marginal product. Feldstein has every right to make this plausible assumption, but it should not be imputed to me, since it is the very question which I set out to investigate. As Feldstein says, I have estimated the parameters of a labor demand curve. However, he simply assumes the demand price in this function includes the payroll tax, whereas I have sought empirical verification.<sup>5</sup> Feldstein may feel this *a priori* proposition is so self-evident as to require no investigation. However, a tradeoff between the employer tax and the nominal wage obtainable for a given amount of labor does not appear to be generally recognized by participants in the U.S. labor market. Neither in the collective bargaining and social security literature, nor in informal discussions, have I encountered the opinion that the

compensation package obtainable is depressed by increases in the employer tax.<sup>6</sup> Hence, I believe my estimates of  $s$  near one are important; this result tells something about the real world, and I do not believe it was forthcoming simply because Feldstein assumed it.<sup>7</sup>

## II. Relevance of the Analysis to "Incidence"

Feldstein states that my estimated labor demand equations (comparable to his relation (5) obtained by assumption) cannot provide any information about tax shifting or the incidence of the tax, because no information is obtained about the response of labor supply or capital stock. What I have shown, he says, is that my data are consistent with CES technology and equality of the gross wage and marginal product of labor. He concedes, however: "At most, his empirical analysis has shown that, *if the supply of labor and capital were not changed by the tax*, the working of the factor markets would cause the entire burden to fall on labor" (p. 738). This is exactly what I claim to have shown by the empirical finding that labor demand price includes the payroll tax. Furthermore, there are reasons to believe that the significance of this finding is greater and more general than suggested by Feldstein.

The generally accepted analysis of tax incidence considers "the resulting change in the distribution of income available for private use" under the assumption of full employment.<sup>8</sup> Under this assumption of a given

<sup>5</sup> Such opinion has surfaced in some European countries where the tax is larger. However, the only tradeoff which appears to be recognized in this country is that between private fringe benefits and the money wage.

<sup>7</sup> So confident is Feldstein that  $s$  must equal one that he feels called upon to explain (p. 737, fn 3) why my estimated values were not identically equal to one. He lists five "econometric problems" to help account for this. However, the list includes factors such as market imperfections which are economic rather than simply statistical. It is factors such as these which prevent many from accepting Feldstein's assumption, and thus make a statistical test worthwhile.

<sup>8</sup> See Richard Musgrave, pp. 208 ff. (Note that my analysis of the impact of the tax on income distribution refers to its effect on the wage share and does not consider the effect on personal income of the distribution of the proceeds by the government.)

<sup>6</sup> In the pioneer study of the CES function, "the data on wage payments for different countries include varying proportions of non-wage benefits," but there is no analysis of the implications of this (see Arrow et al, p. 227).

level of employment, the implication of my empirical analysis is that the payroll tax comes out of the share of labor. However, once variation of factor supplies and employment is admitted, the very concept of the incidence of a broad-based tax becomes murky. Indeed the methodology of tax incidence analysis in this context is largely unexplored. However, I explicitly treated these general equilibrium considerations as a qualification of my findings and argued that the significance of employment effects was likely to be minor. Nevertheless, Feldstein concluded that the absence of empirical analysis of the effect of the tax on factor supplies robs my analysis of almost all relevance to the incidence question. In so doing he chose to ignore the substance of this discussion and chided me on two minor points.

Feldstein first took note of my point that the employer tax would be fully borne by labor if aggregate labor supply were completely inelastic, or the supply price of labor included the full amount of the tax. The impression was created that these extreme conditions are the *sine qua non* of my argument. However, all that is needed for the main burden to fall on labor is that *either* be *approximately* true.<sup>9</sup> I suggested that each of these tendencies was likely and that together they would be reinforcing in support of the labor burden hypothesis. A high degree of inelasticity of aggregate labor supply remains plausible to me, since workers cannot ordinarily withdraw only a part of their labor, cannot hold out altogether for long and have nowhere to hide from a universal tax. It also seems likely that workers do regard at least part of social security contributions as their own saving, in the same way that they are obviously aware of the tradeoff between wages and private pensions negotiated as a package. The lack of any indication of especially high unemployment rates in countries with 10 to 40 percent payroll tax rates is also consistent with the operation of one or both of the above tendencies.

Feldstein also concentrated on one point I made in the support of the possibility that

<sup>9</sup> Actually, as will be shown later, these are sufficient but not necessary conditions.

labor regards much of the employer "contribution" as part of its income. He notes that in this particular context I had implicitly taken for granted that the employee portion of the tax was regarded as saving and included in the supply price of labor. I then suggested that the employer tax would probably also be included in supply price, since there was no reason for labor to regard the two parts of the tax differently. I regret my failure to make this assumption about the employee tax explicit, but I do not feel as though I have been caught dealing from the bottom of the deck. Such an assumption seems quite plausible in the case of this explicit earmarked deduction from the contractual wage—a deduction which wage-earners are encouraged to regard as a "premium," entitling them to later benefits.<sup>10</sup>

Feldstein ignored altogether my *a priori* reasoning (pp. 115–17) which does *not* depend on either the assumption that labor views payroll taxes as part of its income or the assumption of inelastic supply. It utilizes Hicks' analysis of the determinants of the elasticity of derived factor demand. I argued there that, even in the case of elastic labor supply, labor could completely avoid the tax only under extremely implausible elasticity assumptions. On the other hand, under quite believable assumptions (including a contraction of employment), the wage bill would be reduced by the full amount of the tax. In fact, if I may borrow Feldstein's Figure 1, he generously offers a convenient and plausible demonstration of the latter result.<sup>11</sup> Since Feldstein happens to have drawn a demand curve of approximately unitary elasticity in the relevant range, the total compensation of labor (including the tax) is  $L_1W_1(1+t)$  at the

<sup>10</sup> This assumption concerning the employee tax is not essential to any other part of my study which concentrates explicitly on the employer tax. However, see Lester Taylor for a regression result which (if the model is properly specified) may be interpreted as an indication that labor views social security contributions as part of its own saving (and therefore saves less privately by that amount).

<sup>11</sup> Feldstein's demonstration depends, of course, on the result which I obtained empirically and which he assumed—that the demand price includes the payroll tax.

reduced level of employment, or about the same as without the tax ( $L_0W_0$ ). Thus the imposition of the tax leaves a net wage bill ( $L_1W_1$ ), which represents a reduction of the no-tax wage bill by the full amount of the tax.

Whether labor is bearing the entire tax in the previous illustration is a matter of definition. Undoubtedly the assumed cut-back in production due to the tax would also cut the return to capital, and the picture would be somewhat different under the relative share criterion of incidence. However, although the admission of employment variability clouds the incidence concept, if the wage bill falls by the amount of the tax or more, it would be very difficult to maintain that labor bears *less* than the full amount of the tax. Also, it is difficult to conceive how Feldstein could alter the elasticities in his diagram sufficiently to prevent the tax from reducing the wage bill. All in all, the qualification introduced by recognition of possible supply elasticities (as insisted upon by Feldstein) may be rather academic.

I have no quarrel with Feldstein's interesting point that the capital formation effect of the social security program may depress real wages. He also makes the valid points that my empirical and theoretical analyses take no account of the effects of the tax on capital stock or on the relative prices of goods consumed by labor and by the owners of capital, respectively. However, there is no reason to expect any impact of the tax on the rate of return or stock of capital if labor bears the tax without substantial employment effects. In common with other studies, I assumed away the problem of relative prices, or "incidence on the side of uses of income," but I do not see this as a troubling assumption.

One does not need to assume that labor and capitalists have the same consumption patterns; it is only necessary that the tax have about the same overall effect on the two sets of prices, which probably amounts to assuming about the same degree of labor intensity in the production of each market basket. (Is there any a priori reason to surmise that the tax will raise the price of yachts more or less than the price of bowling balls?)

In sum, my empirical finding that the payroll tax comes out of the real demand price of labor has considerable significance if one accepts my a priori suggestions that employment effects are probably minor. Then labor clearly bears the tax. Even if employment is reduced by the tax, there appears to be no reason to expect that the wage bill will be reduced by less, rather than more, than the amount of tax. All of this may seem like beating a dead horse to some theorists who feel the labor burden thesis follows directly as a corollary of established theory. However, Feldstein's reaction, along with continuing skepticism among labor and social security specialists, suggests that the issue may not be closed.

#### REFERENCES

- K. Arrow, H. B. Chenery, B. S. Minhas, R. M. Solow, "Capital-Labor Substitution and Economic Efficiency," *Rev. Econ. Statist.*, Aug. 1961, 43, 225-50.
- L. D. Taylor, "The Marginal Propensity to Save Out of Different Types of Income," in A. M. Okun and G. L. Perry, eds., *Brookings Papers on Economic Activity*, 2:71 Washington 1971.
- R. Musgrave, *The Theory of Public Finance*, New York 1959.

# Decision Rules for Effective Protection in Less Developed Economies

By TRENT J. BERTRAND\*

It is the purpose of this note to show that the concept of effective rate of protection is useful in defining protection policies given the constraints under which policy is often formulated in less developed economies. We study the appropriate structure of protection when it is desired that a given amount of value-added be generated in a high priority sector of the economy.<sup>1</sup> Arguments for giving *general* preferences to industrial activity, for instance, are often based on alleged dynamic affects on technology, the quality of the labor input, and the quality of entrepreneurship.<sup>2</sup> Whether or not it is certain that these types of social benefits are specific to industry in general or certain types of industry in particular, they clearly are felt to be by policy makers in less developed countries. This paper is aimed at developing rational guidelines to policies based on these considerations.

## I. Decision Rules for Protection: Basic Results

The problem considered here is to maximize the consumption possibilities open to the domestic economy subject to the constraint

\* Associate professor of political economy at The Johns Hopkins University. I am indebted to Bela Balassa, George Borts, Frank Flatters and a referee for comments on an earlier draft of this paper.

<sup>1</sup> The usefulness of the concept of effective protection has been criticized in several recent contributions (see Roy Ruffin, A. H. H. Tan (1970), V. K. Ramaswami and T. N. Srinivasan, Bertrand and Vanek (1972). These criticisms are based on difficulties in using effective rates to predict resource movements between industries but are not relevant to the results established here that indicate resource movements, whatever their specific nature, will be optimal given the policy constraint. For an independent discussion of noneconomic objectives in the framework of a two final product, two primary factor model, see related papers by Tan (1971) and Jagdish Bhagwati and Srinivasan.

<sup>2</sup> These and related arguments for giving preference to industry are discussed in I. M. D. Little, Tibor Scitovsky and Maurice Scott and Bela Balassa.

that a certain amount of value-added is generated in the industrial sector.<sup>3</sup> Consumption possibilities are equal by the budget constraint to gross value of output minus the value of intermediate inputs; that is,

$$(1) \quad \sum_i P_i C_i = \sum_i P_i x_i - \sum_i \sum_j P_j a_{ji} X_i$$

$$i, j = 1 \dots n$$

where  $P_i$  is the invariant international price of the  $i$ th commodity,  $C_i$  is consumption of the  $i$ th commodity,  $x_i$  is the output of the  $i$ th commodity, and  $a_{ji}$  is the constant amount of  $j$ th commodity used in the production of one unit of the  $i$ th commodity. Denoting value-added per unit in the production process of the  $i$ th commodity as  $\pi_i$  and substituting relation (2) into relation (1), we obtain relation (3) showing that the value of the consumption bundle is equal to total domestic value-added.

$$(2) \quad \pi_i = P_i - \sum_j a_{ji} P_j$$

$$(3) \quad \sum_i P_i C_i = \sum_i \pi_i x_i$$

The problem is therefore to maximize total domestic value-added subject to the constraint on production defined by a strictly convex gross production possibility set  $T$  and the policy constraint on required value-added generated in the industrial sector; that is,

$$(4) \quad \text{Max} \sum_i \pi_i x_i \quad i = 1 \dots r, r + 1 \dots n$$

subject to

$$(5) \quad T(x_1 \dots x_n) = 0$$

$$i = 1 \dots r, r + 1 \dots n$$

<sup>3</sup> The relationship of maximizing consumption possibilities to maximizing social welfare is discussed in Section III below.

and subject to

$$(6) \quad \sum_h \pi_h x_h \geq K \quad h = 1 \dots r$$

where  $h$  denotes the priority industrial activities and  $K$  is an exogenously determined constraint.

Forming the Lagrangean expression (7), taking partial differentials

$$(7) \quad Z = \sum_i \pi_i x_i - \lambda T(x_1 \dots x_n) - \phi \left( \sum_h \pi_h x_h - K \right)$$

with respect to changes in outputs, and setting these equal to zero, we obtain the first-order maximum conditions for the nonindustrial activities (denoted by subscript  $k$ ,  $n-r$  in number);<sup>4</sup>

$$(8) \quad \frac{\partial Z}{\partial x_k} = \pi_k - \lambda T_k = 0$$

and for the  $r$  industrial activities (denoted by  $h$  subscript);

$$(9) \quad \frac{\partial Z}{\partial x_h} = \pi_h - \lambda T_h - \phi \pi_h$$

<sup>4</sup> The conditions given in relations (8)–(9) are the first-order conditions for a maximum. The satisfaction of the second-order conditions is guaranteed by the assumption of strict convexity of the gross production possibility set  $T$ . As shown by Bradley Billings, this requires restrictions on the nature of factor substitution and requires that the factor intensity of any output is not a proportional linear combination of the factor intensities of the remaining goods. The latter requirement implies that there are at least as many products as factors, or otherwise the production sets have linear segments as shown by, for instance, Bertrand and Vanek (1971a). Even with an equal or greater number of products than factors, the condition may be violated but such a case may legitimately be disregarded as a perverse case. The restrictions on input substitution require that equal product contours or isoquants, when projected onto the factor and intermediate input planes, yield convex to the origin curves. This is a weaker restriction than imposed in this paper, where it is assumed that intermediate inputs are used in fixed proportions. Our stronger restriction is required for maintaining the equality between marginal rates of transformation and relative unit values added (as given in relation 13) and not for strict convexity of  $T$ . On this latter point, see Ethier 1970.

Thus, for maximum consumption possibilities, we have for any two nonindustrial activities (say  $k=l$  and  $k=m$ );

$$(10) \quad \frac{\pi_l}{\pi_m} = \frac{T_l}{T_m}$$

and we have for any two industrial activities (say  $h=g$  and  $h=s$ );

$$(11) \quad \frac{\pi_g}{\pi_s} = \frac{T_g/(1-\phi)}{T_s/(1-\phi)} = \frac{T_g}{T_s}$$

and we have for any combination of industrial and nonindustrial activities (say  $k=l$  and  $h=g$ );

$$(12) \quad \frac{\pi_l}{\pi_g} = \frac{T_l}{T_g/(1-\phi)}$$

With the normal neoclassical assumptions, perfect competition will lead to production where the marginal rate of transformation along the gross production possibility set is equated to the domestic unit values-added;<sup>5</sup> that is,

$$(13) \quad \frac{T_i}{T_j} = \frac{\pi_i(1+z_i)}{\pi_j(1+z_j)}$$

where  $z$  is the effective rate of protection, the percentage distortion between free trade and protected value-added.

Relations (10)–(13) are basic to our analysis. The following decision rules for effective protection are immediately established:

- A: The effective rate of protection should be equalized on all nonindustrial activities (from (10) and (13)).
- B: The effective rate of protection should be equalized on all industrial activities (from (11) and (13)).
- C: There should be a common rate of effective discrimination between industrial and nonindustrial activities (from (12) and (13) and implied by A and B).

These decision rules underlie the relevance of the concept of effective tariffs for designing the structure of protection in less developed countries. While following the equal dis-

<sup>5</sup> This is proved in Bertrand and Vanek (1971a).

crimination in favor of industrial activities guidelines can be expected to lead to non-uniform changes in production among industries compared to the free trade situation, this distortion in resource allocation will be optimal.

## II. Decision Rules for Protection: Some Extensions

The decision rules derived in Section I for the situation where preferential treatment is given the industrial sector so as to attain a fixed amount of industrial activity can be extended in several ways.

First, if the beneficial effects of industrial activity are not uniform for all types of industrial activity, it may still be feasible to assign different priorities to groups of industries. Thus, the policy constraint of Section I could be reformulated as (14)

$$(14) \quad \beta^1 \sum_w \pi_w x_w + \beta^2 \sum_y \pi_y x_y + \dots + \beta^v \sum_z \pi_z x_z \geq K$$

where  $\beta^1, \beta^2, \dots, \beta^v$  are the weights reflecting the priority attached to the  $v$  groups of industries. Substituting (14) for (6) and solving for the first-order conditions for a maximum, it is seen that relative unit values-added should be equated to marginal rates of transformation (i.e., effective protection should be equalized within industry groups) while for any two industries in different groups (say industry  $c$  in group 2 and industry  $e$  in group  $v$ ), we have;

$$(15) \quad \frac{\pi_c}{\pi_e} = \frac{T_c/(1 - \phi\beta^2)}{T_e/(1 - \phi\beta^v)}$$

which again defines the uniform degree of discrimination between sectors.

Second, it might be more appropriate to design the policy constraint so that a certain percentage of value-added must be generated in the industrial sector rather than an absolute amount. The constraint (6) would then be replaced by (16),

$$(16) \quad \sum_h \pi_h x_h \geq \alpha \sum_i \pi_i x_i$$

where  $\alpha$  is the proportion of total domestic value-added (generated by all  $i$  activities and evaluated in world prices) that must be generated in the industrial sector (where industrial activities are again denoted by the  $h$  subscript). By again forming the Lagrangean with (16) substituted for (6) and solving for the first-order conditions for a maximum, we confirm that equivalent rates of effective protection are required between industrial and nonindustrial activities and that a uniform degree of discrimination, defined by (17), is required between the two sectors;

$$(17) \quad \frac{\pi_c}{\pi_e} = \frac{T_c/(1 + \phi\alpha)}{T_e/(1 + \alpha\phi - \phi)}$$

## III. Decision Rules for Protection: Qualifications

The policies considered in this paper are designed to maximize the consumption possibilities open to a small economy subject to several alternative constraints. However, tariffs also distort consumption so that it is possible that maximum consumption possibilities do not correspond to maximum welfare. The analysis could be reworked for effective rates of subsidy which would not affect the prices facing consumers, and the results would then not have to be qualified by reference to possible consumption losses. However, the revenue constraint dictates strongly in favor of using tariffs rather than subsidies in the viewpoint of policy makers in most less developed economies. It therefore appears more realistic to deal with tariffs and to note that it is possible that consumption distortions could make a maximum consumption possibilities policy nonoptimal. A system of consumption taxes would then be required to eliminate consumption distortions.

Another qualification is required by our assumption of the "small country" case. Certain less developed economies may have significant world market shares for selected traditional export products and therefore some potential to influence world prices for these goods. However, this does not greatly affect our results even though they are derived for industries in which world product

and input prices are constant. First, fixed price conditions are likely to hold for the industries to which policy makers tend to give priority. Thus, the equivalent effective rates of protection result for these industries is not affected since relations (9) and (11) still hold. Second, the uniform rate of distortion in protection between industries will still hold for priority and nonpriority industries other than the few industries where market power exists as seen by relations (8), (9), and (12). The degree of market power would, however, have to be taken into account in defining the optimal rates of protection to traditional export industries wherever the constant price assumption is not valid.

Finally, it might be noted that many reasons may exist for giving differential treatment to industries in less developed economies. Thus, if the wage rate facing the industrial sector overstates the opportunity cost of labor, as is often argued, and subsidies related to wage payments are not feasible, then protection may be justified. The optimum can be expected to involve higher rates of effective protection to labor intensive activities. Similarly, where external economies are present, they may differ significantly between industries, as could the training effects underlying the dynamic considerations used as a justification for tariff protection in Section I. If savings rates are related to the returns to capital, differential rates in favor of capital intensive industries might be justified. Finally, the interdependence between sectors based on the inducement mechanisms identified in Hirschman's work also might justify differential treatment. The decision rules derived in Section I, however, do provide general guidelines for protection policy within which exceptions might be granted should special social benefits be identified with particular activities. This would replace procedures presently used in many less developed countries where protection is tailor-made to make possible a specified rate of return in all protected investment. Section II has also shown that where differential protection is called for, if it is possible to classify industries into several groups within which the distortion between social and private

profitability is uniform, our decision rules remain applicable.

## REFERENCES

- T. J. Bertrand and J. Vanek, "Effective Protection and Resource Allocation" in W. Sellekarts, ed., *Papers in Honor of Jan Tinbergen*, forthcoming 1972.
- and ——— (1971a), "Trade and Factor Prices in a Multi-Commodity World" in J. N. Bhagwati, et al., eds., *Trade, Balance of Payments and Growth*, Amsterdam 1971.
- and ——— (1971b), "The Theory of Tariffs, Taxes, and Subsidies: Some Aspects of the Second Best," *Amer. Econ. Rev.*, Dec. 1971, 61, 925-31.
- B. Balassa, *The Structure of Protection in Developing Countries*, Baltimore 1971.
- J. Bhagwati and T. N. Srinivasan, "Optimal Intervention to Achieve Non-Economic Objectives," *Rev. Econ. Stud.*, Jan. 1969, 35, 27-38.
- B. Billings, "General Equilibrium Production Effects of Effective Rates of Protection," unpublished doctoral dissertation, Cornell Univ. 1971.
- W. J. Ethier, "General Equilibrium Theory and the Concept of the Effective Rate of Protection," disc. paper No. 170, Univ. Pennsylvania, June 1970.
- I. M. D. Little, T. Scitovsky, and M. Scott, *Industry and Trade in Some Developing Countries, A Comparative Study*, London 1970.
- V. K. Ramaswami, and T. N. Srinivasan, "Tariff Structure and Resource Allocation in the Presence of Factor Substitution," in J. N. Bhagwati et al., eds., *Trade, Balance of Payments and Growth*, Amsterdam 1971.
- R. Ruffin, "Tariffs, Intermediate Goods, and Domestic Protection," *Amer. Econ. Rev.*, June, 1969, 59, 261-69.
- A. H. H. Tan, "Differential Tariffs, Negative Value-Added and the Theory of Effective Protection," *Amer. Econ. Dev.*, Mar. 1970, 60, 107-16.
- , "Optimal Trade Policies and Non-economic Objectives in Models Involving Imported Material, Inter-Industry Flows, Non-traded Goods," *Rev. Econ. Stud.*, Jan. 1971, 38, 105-12.

# Substitution, Complementarity, and the Residual Variation: Some Further Results

By LOUIS PHILIPS AND PHILIPPE ROUZIER\*

An earlier paper by Philips reported the results of a principal component analysis of the residual correlation matrix obtained after estimating a system of dynamic demand equations. The purpose was to verify the postulated additive nature of the utility function and to collect some information on possible substitution and complementarity relationships among the commodity groups. The residuals were obtained using the original Houthakker-Taylor (HT) 1966 estimation procedure. This procedure presents some advantages, but also some deficiencies. On the other hand, the correlations (in particular their signs) were interpreted on the basis of the Hicksian definitions of substitutability and complementarity, in terms of the signs of the substitution effects.

The object of this note is to present some further results. First, it is of some interest to determine to what extent the results reported in the abovementioned article resist not unimportant changes in the estimation procedure. Secondly, the Hicksian definitions are rather deceptive: it is intuitively more appealing to work with the old (cardinal) notions of substitutability and complementarity (stated in terms of the signs of the second cross partial derivatives of the utility function). A decomposition of the (total) residual correlations into "general" and "specific" correlations, the latter corresponding to preference relations defined in cardinal terms, is presented here. This corresponds to a breakdown of the substitution effect into a general and a specific effect. The analysis of this specific effect allows us to check our previous conclusions as to the grouping of commodi-

ties for which the assumption of additive preferences is appropriate.

## I. Estimation of the Model

We first consider the estimation procedure. To begin with, it should be noticed that the budget constraint of the consumer has to be defined in terms of undeflated prices (and therefore undeflated "income"). In the process of estimation, this constraint should therefore be measured as

$$(1) \quad \sum_i p_{it} \hat{q}_{it} = y_t \quad (i = 1, \dots, n)$$

where  $p_{it}$  is the nominal price, i.e., the implicit price deflator for commodity  $i$ ;  $\hat{q}_{it}$  is the estimated quantity, i.e., the estimated expenditure in constant prices; and  $y_t$  is observed income or total expenditures in current prices. Condition (1) is not the same as the condition that total estimated expenditures in constant prices be equal to total observed expenditures in constant prices.

Introducing this correction, one gets a new formula for estimating  $\lambda_t$  after substituting into (1) the estimating equations

$$(2) \quad q_{it} = K_{i0} + K_{i1}q_{i(t-1)} + K_{i2}\lambda_t p_{it} + K_{i3}\lambda_{t-1}p_{i(t-1)}$$

derived by HT from a generalized quadratic utility function. Indeed, one obtains equation (3). Notice that  $\lambda_t$  is now homogeneous of degree minus one in  $p_{it}$  and  $y_t$ , where  $p_{it}$  is the undeflated price and  $y_t$  is total expenditures in current prices. There is no need therefore to use relative prices in (2), i.e., to deflate  $p_{it}$  by the deflator of total expenditures. Strangely enough, HT continue, after introducing the correct budget constraint (1), to use relative prices in the estimations of (2) presented in chapter 5 of the second edition of *Consumer Demand in the United States* (p.

\* Professor of economics at the Catholic University of Louvain and research assistant at the University of Michigan, respectively. Thanks are due to Anton Barten and C. Lluch for stimulating discussions and to J. P. Lemaître and R. Sanz-Ferrer for programming and research assistance.

$$(3) \quad \lambda_t = \frac{y_t - \sum \hat{K}_{i0} p_{it} - \sum \hat{K}_{i1} p_{it} q_{i(t-1)} - \lambda_{t-1} \sum \hat{K}_{i2} p_{it} p_{i(t-1)}}{\sum \hat{K}_{i2} p_{it}^2}$$

201). Clearly, this introduces unnecessary complications.

One could also quarrel HT's (and our) use of "three-pass" least squares, given that Kenneth Wallis has shown that this method does not lead to consistent estimates when the variables other than the lagged dependent variable are autocorrelated. The results reported here are obtained by using ordinary least squares, which amounts to ignoring the problem of autocorrelation. Some justification, other than simplicity of computation, might be given in terms of an argument developed by Robert A. Pollak and Terence J. Wales. The argument says that the introduction of past decisions in the model takes care of precisely that relationship which typically creates autocorrelation of the errors.

Finally, there is the important fact, emphasized by Richard W. Parks and others, that the budget constraint implies

$$(4) \quad \sum_i p_{it} \epsilon_{it} = 0$$

which in turn implies that the contemporaneous variance-covariance matrix is singular and not constant over time, given the presence of  $p_{it}$  in (4). To make the hypothesis of a constant variance-covariance matrix plausible, one has therefore to multiply (2) by  $p_{it}$  on the left- and the right-hand side. To get efficient estimators, one would have to use for example, Arnold Zellner's method for seemingly unrelated regressions using a generalized inverse (see J. Plasmans). The first step in this method is typically to proceed equation by equation, using ordinary least squares (to get an estimate of the variance-covariance matrix). The present use of ordinary least squares may be interpreted as corresponding to that first step.

## II. Total and Specific Correlations

Under the assumptions given in Philips it can be shown that the residual correlation matrix gives valuable information on the structure of preferences. If all correlations turn out to be negative, we are not allowed

to reject the hypothesis that the (quadratic) utility function is additive. If the signs differ, we may interpret a negative correlation to indicate that goods  $i$  and  $j$  are substitutes in the Hicksian sense, and a positive that they are complements in the Hicksian sense. In other words, these signs are the opposite of the signs of the (total) cross substitution effects.

Notice that the above test is biased in favor of additivity. Indeed, as  $\epsilon_t = K^* \Delta a_t$ , where  $K^*$  is proportional to the Slutsky matrix  $K$ , and  $\Delta a_t$  is a vector of random shocks affecting the utility function, the correlation matrix  $R$  is equal to  $K^* E[(\Delta a_t)(\Delta a_t)'] K^*$  after normalization. As the homogeneity condition implies  $K^* \cdot p = 0$ ,  $R \cdot p = 0$ . Therefore  $R$  is singular, so that at least one correlation must be negative in each row of  $R$ . If nevertheless some correlations turn out to be positive, the point that this indicates absence of additivity is reinforced. Conversely, the singularity of  $R$  weakens the interpretation to be given to the signs in terms of substitutability and complementarity.

It is more important, therefore, to find a way to shift from Hicksian concepts to cardinal concepts of substitutability and complementarity. Such a procedure was suggested to us in private correspondence by Anton P. Barten. The idea is to subtract the correlation due to the general substitution effect (see Houthakker) from the "total" correlation coefficients to get the correlation due to the specific substitution effect. (This is, after all, what we were implicitly looking for, as the specific effect informs directly about the structure of preferences.) The argument runs as follows.

Under independence,<sup>1</sup> the covariance be-

<sup>1</sup> Notice that Barten has shown that, under want independence,

$$k_{ij} = \phi \frac{\delta q_i}{\delta y} \frac{\delta q_j}{\delta y},$$

where  $\phi$  is a constant of proportionality and  $k_{ij}$  is the (total) cross substitution effect.

tween  $\epsilon_i$  and  $\epsilon_j$  is

$$(5) \quad -\kappa \frac{\delta q_i}{\delta y} \frac{\delta q_j}{\delta y} \quad i \neq j \quad \kappa > 0$$

while the variance of  $\epsilon_i$  is

$$(6) \quad \kappa \frac{1}{p_i} \frac{\delta q_i}{\delta y} \left(1 - p_i \frac{\delta q_i}{\delta y}\right)$$

The correlation between  $\epsilon_i$  and  $\epsilon_j$  is therefore, under independence,

$$(7) \quad -\sqrt{\frac{p_i \frac{\delta q_i}{\delta y} p_j \frac{\delta q_j}{\delta y}}{\left(1 - p_i \frac{\delta q_i}{\delta y}\right) \left(1 - p_j \frac{\delta q_j}{\delta y}\right)}}$$

Using estimates of the income derivatives one can compute (7) and subtract it from the corresponding total correlation. This amounts to adding a positive number to each total correlation coefficient, and therefore to increasing specific complementarity. A principal component analysis of the resulting specific correlations should lead to a grouping criterion that gives a clearer idea of the preference relations among commodities.

Reestimating the HT model, for the eleven commodity series of the *Survey of Current Business* over the period 1949-1969, after introducing the corrections listed in the preceding section, the following results shown in Tables 1 and 2 were obtained after 23 iterations.

The matrix of *total* correlations (Table 1)

TABLE 1—MATRIX OF TOTAL CORRELATIONS (R), ELEVEN VARIABLES (LAST-STEP RESIDUALS)

	1	2	3	4	5	6	7	8	9	10
1. Auto & Parts	1.000									
2. Furniture & Equipment	-0.345	1.000								
3. Other Durables	-0.629	0.366	1.000							
4. Food & Beverages	-0.649	-0.287	0.252	1.000						
5. Clothing & Shoes	-0.475	0.305	0.805	-0.076	1.000					
6. Gasoline & Oil	0.451	-0.155	-0.034	-0.553	0.193	1.000				
7. Other Nondurables	-0.453	0.326	0.195	0.005	0.470	-0.029	1.000			
8. Housing	-0.041	0.164	-0.010	-0.338	0.206	0.314	0.049	1.000		
9. Household Operation	0.365	0.149	-0.238	-0.564	-0.005	0.361	-0.086	0.060	1.000	
10. Transportation	0.347	-0.439	-0.240	0.037	-0.354	0.069	-0.417	-0.206	-0.008	1.000
11. Other Services	-0.511	0.066	-0.070	0.500	-0.279	-0.688	-0.035	-0.147	-0.348	-0.177

TABLE 2—MATRIX OF SPECIFIC CORRELATIONS

	1	2	3	4	5	6	7	8	9	10
1. Auto & Parts	1.000									
2. Furniture & Equipment	-0.018	1.000								
3. Other Durables	-0.500	0.434	1.000							
4. Food & Beverages	-0.266	-0.086	0.331	1.000						
5. Clothing & Shoes	-0.222	0.438	0.857	0.080	1.000					
6. Gasoline & Oil	0.519	-0.120	-0.020	-0.511	0.220	1.000				
7. Other Nondurables	-0.264	0.425	0.234	0.121	0.547	-0.009	1.000			
8. Housing	0.046	0.209	0.008	-0.285	0.241	0.323	0.075	1.000		
9. Household Operation	0.462	0.200	-0.218	-0.504	0.034	0.372	-0.057	0.074	1.000	
10. Transportation	0.436	-0.392	-0.222	0.092	-0.318	0.079	-0.390	-0.194	0.006	1.000
11. Other Services	-0.313	0.170	-0.029	0.622	-0.199	-0.667	0.025	-0.119	-0.317	-0.149

has basically the same properties as in Philips. Again almost half (23 out of 55) are positive. Furthermore, those (Hicksian) substitution and complementarity relationships that were very pronounced remain unchanged. Cars are substitutes for all other durables; clothing is a complement to other durables (jewelry, watches, ophthalmic products, books, toys, etc.); food appears as a substitute for housing and household operation.

But there are a number of differences, some of which are not unimportant and may be considered as improvements. For example, gasoline is now clearly complementary to automobiles; "other services" (including radio and TV, medical care, higher education etc.) are substitutes for cars and complementary to food and beverages.

Our main interest goes to Table 2, however. There we find indications of *specific* (cardinal) preference relationships, such as complementarity of cars, gasoline, household operation and transportation on the one hand, and of clothing and other durables on the other hand; substitutability of food and expenses for household operation; complementarity of clothing and the item "other nondurables," which includes tobacco, toilet articles, cleaning, stationary, flowers etc.; substitutability between gasoline and other services and a more pronounced complementarity of food and "other services." All this tends to the conclusion that the improvements in the estimation of the model

led to better results in terms of residual correlations.

What sort of groupings come out of these *specific* relationships? Table 3 presents the factor loadings ( $F_1$ ) for the first six principal components, accounting for 89.8 percent of the total variance. (We thus cut off four columns of (very low) factor loadings, the last eigenvalue being zero.) These components were rotated by the varimax procedure, to get a clearer idea of the grouping (which is what interests us here).

The grouping is very clear. The six groups include respectively: Furniture and household equipment and household operation; Clothing and "other durables"; Food and beverages, gasoline and oil and "other services"; Automobiles and parts and transportation; Housing; "Other nondurables." These groups do seem to correspond to basic needs.

Is this a partition of the commodity set such that the utility function is additive? The philosophy of our approach suggests to use two tests (which should lead to non-conflicting answers). One test is to rotate the six components of Table 3 further and to see whether the matrix of intercorrelations among the resulting oblique factors is unitary. The other test is to rearrange the data in six groups according to the partition given in the preceding paragraph, to re-estimate the model (with  $n=6$ ) and to see whether all residual total correlations have a negative sign.

TABLE 3—FIRST-ORDER FACTOR LOADINGS ( $F_1$ ), AFTER VARIMAX ROTATION

	1	2	3	4	5	6
1. Auto & Parts	0.385	-0.116	0.278	0.757	0.106	-0.002
2. Furniture & Household Equipment	0.701	0.410	-0.261	-0.172	0.214	0.265
3. Other Durables	-0.049	0.591	-0.068	-0.172	-0.034	0.012
4. Food & Beverages	-0.396	0.280	-0.734	0.232	-0.178	0.150
5. Clothing & Shoes	0.087	0.862	0.176	-0.090	0.143	0.372
6. Gasoline & Oil	0.065	0.069	0.797	0.323	0.282	0.102
7. Other Nondurables	0.043	0.207	-0.024	-0.198	0.009	0.239
8. Housing	0.041	0.055	0.152	-0.070	0.962	0.019
9. Household Operation	0.776	-0.126	0.392	0.133	-0.066	-0.047
10. Transportation	-0.199	-0.090	0.031	0.789	-0.182	-0.310
11. Other Services	0.002	-0.112	-0.927	-0.103	0.048	0.025

Interestingly, both tests are negative. The eigenvalues of the matrix of intercorrelations (among the six oblique components) are respectively: 1.472, 1.412, 0.982, 0.774, 0.742, and 0.617. Of the fifteen total residual correlations of the system aggregated into the same six commodity groups, eight only have a negative sign.

To proceed further, one could factorize the matrix of intercorrelations (among the six oblique factors) and apply the same tests to partitions in a smaller number of groups. All attempts in that direction gave indications pointing towards the absence of additivity.

Our conclusion remains unaltered: given the eleven consumption categories of the *Survey of Current Business*, the assumption of additive preferences cannot be accepted at that level of aggregation nor at higher levels.

#### REFERENCES

- A. P. Barten, "Preference and Demand Interactions between Commodities," CORE disc. paper 7027, Louvain, Aug. 1970.
- H. S. Houthakker, "Additive Preferences," *Econometrica*, Apr. 1960, 28, 244-57.
- and L. D. Taylor, *Consumer Demand in the United States, 1929-1970*, 2d ed., Cambridge, Mass. 1970.
- and ———, "Joint Estimation of Dynamic Demand Equations," mimeo 1966.
- R. W. Parks, "Systems of Demand Equations: An Empirical Comparison of Alternative Functional Forms," *Econometrica*, Oct. 1969, 37, 629-50.
- L. Philips, "Substitution, Complementarity, and the Residual Variation Around Dynamic Demand Equations," *Amer. Econ. Rev.*, Sept. 1971, 61, 586-97.
- J. Plasmans, *The General Seemingly Unrelated Regression Problem*, E.I.T. Nos. 18 and 19, Tilburg 1970.
- R. A. Pollak and T. J. Wales, "Estimation of the Linear Expenditure System," *Econometrica*, Oct. 1969, 37, 611-28.
- K. F. Wallis, "Lagged Dependent Variables and Serially Correlated Errors: A Reappraisal of Three-Pass Least Squares," *Rev. Econ. Statist.*, Nov. 1967, 49, 555-67.
- A. Zellner, "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias," *J. Amer. Statist. Ass.*, June 1962, 57, 348-68.
- U.S. Office of Business Economics, *Survey of Current Business*, various issues, 1949-69.

# Schooling and Earnings of Low Achievers: Comment

By BARRY R. CHISWICK\*

In their article in this *Review*, W. Lee Hansen, Burton Weisbrod, and William Scanlon (HWS) conclude that "the estimated payoff to more schooling is low" for their sample of men who were rejected for military service due to low scores on the Armed Forces Qualification Test (AFQT). It is shown in this comment that they obtained a downward biased estimate of the true profitability of schooling for their sample and that the correct rate of return may be at least 13 percent. While this is lower than the rate of return to males from high school, it is similar to the private return from college.<sup>1</sup>

The HWS data are a sample taken in November 1963 of approximately 2,400 males, aged 17 to 25. The dependent variable ( $Y$ ) is annual income in 1962 after deducting transfer payments. The explanatory variables include years of schooling ( $S$ ), age ( $A$ ), AFQT score ( $AFQT$ ), and a dummy variable which takes the value of one if the individual received "training outside of school" ( $T$ ). Their model is

$$(1) \quad Y = b_0 + b_1S + b_2A + b_3AFQT + b_4T$$

neglecting the other variables for this discussion.<sup>2</sup>

HWS measure the effect of schooling holding age constant. However, age reflects years of schooling and years of labor market experience. By holding age constant, an additional year of schooling means a year less of experience. Experience includes information about the nature and wage structure of jobs, as well as knowledge which directly in-

creases productivity on the job. Recent studies indicate that experience is more important than age for explaining earnings.<sup>3</sup> In addition the positive effect of experience on earnings is strongest for young workers, and the HWS sample is composed of young males. Thus, if age is held constant the slope coefficient of schooling is biased downward because the positive effect on income of an additional year of schooling is partially offset by the reduction in experience.

Estimates of the effect of schooling on earnings holding experience constant can be obtained from the data presented by HWS if it is assumed age has no effect independent of schooling and experience, and that experience ( $E$ ) is measured as age minus schooling minus 5. Substituting  $A = S + E + 5$  into equation (1) results in

$$(2) \quad Y = (b_0 + 5b_2) + (b_1 + b_2)S + (b_2)E + b_3AFQT + b_4T,$$

and the direct effect of schooling on earnings is the sum of the slope coefficients of schooling and age from equation (1).

Rows 1 through 5 in Table 1 present for several sets of control variables, the HWS estimates of the effect of schooling ( $b_1$ ) and my estimates ( $b_1 + b_2$ ) based on the assumption that age has no separate effect. The HWS estimates imply that an extra year of schooling is not profitable, whereas the second set implies that it is profitable.<sup>4</sup>

\* See Jacob Mincer, N. Arnold Tolles, Alice Hanson, and Ewan Clague; and Tolles and Emanuel Melichan. For example, in a study of economists' salaries using 1964 data, age "was found to have only minor net effect after the length of experience was taken into account" (see Tolles et al., p. 65) and a similar conclusion emerged from 1966 data (see Tolles and Melichan, p. 61).

<sup>4</sup> If it is assumed that 100K percent of a full year's potential earnings are invested in schooling,  $Y_s = Y_0(1+rK)^s$ , where  $Y_s$  is earnings after  $S$  years of schooling and  $r$  is the internal rate of return. (See Becker and Chiswick.)  $\ln Y_s = \ln Y_0 + S \ln(1+rK) = \ln Y_0 + rKS$ , where  $rK$  is small, and  $r = (1/YK)(\partial Y/\partial S)$ . It is assumed here that  $K = 1.0$  and earnings are evaluated at

\* Queens College and Graduate Center, City University of New York, and National Bureau of Economic Research. I wish to thank Jacob Mincer for his comments.

<sup>1</sup> See Gary Becker, part II, and Hansen. The mean and standard deviation of years of schooling in this sample are 8.94 and 2.40 years, respectively. (HWS, p. 418).

<sup>2</sup> These are dummy variables for color, region, marital status of the individual and divorce status of his parents, and a variable for family size during childhood.

TABLE 1—EFFECT OF SCHOOLING ON EARNINGS

Row	HWS Model	Age Constant		Experience Constant	
		Direct Effect <sup>a</sup> (\$ per year)	Internal Rate of Return (percent) <sup>c</sup>	Direct Effect <sup>b</sup> (\$ per year)	Internal Rate of Return (percent) <sup>c</sup>
1	I	61.5	3.5	252.6	14.2
2	II	30.3	1.7	232.0	13.1
3	III	25.7	1.5	224.7	12.7
4	III'	55.9	3.2	244.5	13.8
5	IV	20.3	1.1	203.3	11.4
6		43.7 <sup>d</sup>	2.5	228.0 <sup>d</sup>	12.8

Source: Columns (1) and (3) from HWS, Table 1, p. 413.

<sup>a</sup> Slope coefficient of schooling ( $b_1$ ).

<sup>b</sup> Sum of slope coefficients of schooling and age ( $b_1 + b_2$ ).

<sup>c</sup> Internal rate =  $\frac{\text{direct effect}}{\text{average income}}$  (see footnote 4 in the text).

<sup>d</sup> Direct effect of schooling plus indirect effect through  $AFQT$ , (based on model IV and HWS [p. 415]) where  $\partial AFQT/\partial S = 1.0$  and  $b_2 = 23.7$ .

For the model (IV) in which the  $AFQT$ , training, race, region, and family background variables are held constant, the rate of return from schooling, if age has no separate effect, is 11.4 percent.

The variables  $AFQT$  and  $T$  were held constant in the regression model IV. However, the test and the training programs followed the cessation of schooling, and the values of the variables  $AFQT$  and  $T$  are both causally related to  $S$ .<sup>5</sup> By holding these variables constant in a regression, downward biased estimates of the profitability of schooling are obtained.

The direct and indirect (via  $AFQT$ ) effects of schooling on earnings are measured by

$$(3) \quad \frac{\partial y}{\partial S} = (b_1 + b_2) + b_3 \frac{\partial AFQT}{\partial S}$$

The slope coefficients  $b_1$ ,  $b_2$ , and  $b_3$  can be obtained from the HWS paper, Table 1, p. 413. The partial slope coefficient  $\partial AFQT/\partial S$

the mean. Therefore,  $r = (1/\$1,776.90)(\partial Y/\partial S)$ . (Average income from Hansen, Weisbrod and Scanlon, p. 418). If the actual earnings of low achievers while they are students exceeds the direct cost to them of their schooling,  $K$  is less than 1.0, and the rates of return computed in this comment are downward biased.

<sup>5</sup> HWS cite a study of the  $AFQT$  which indicates that the examinee's score depends on several factors, including his level of schooling, p. 411. The simple correlation between  $S$  and  $T$  is significantly positive (their Table A).

was estimated by HWS from a regression of  $AFQT$  on  $S$  and the other independent variables, and was found to be equal to one point of  $AFQT$  per year of schooling, p. 415.<sup>6</sup>

Holding race, region, training, and family background variables constant (HWS model IV) the effect of schooling on income is presented in row 6 of Table 1. Adding the effect of schooling on income via  $AFQT$  to the direct effect of schooling raises the rate of return from schooling from 11.4 to 12.8 percent.

It seems reasonable to assume that for most of the sample the training program ( $T=1$ ) has a positive net present value of income. If, as also seems plausible, greater schooling increases the probability of being selected for a training program, or increases the net present value of the program by reducing entry costs, part of the effect of greater training on income is attributable to schooling. Ignoring this "option value" of schooling most likely generates a small downward bias in the effect of schooling on income.<sup>7</sup>

Therefore, the conclusion of a regression analysis of the profitability of schooling for the HWS sample of low achievers (or any sample) depends on whether it is age or years of labor market experience which is held constant. When age is held constant the rate of return from schooling is biased downward. When experience is held constant, the rate of return from schooling is at least 13 percent for low achievers.

## REFERENCES

- G. S. Becker, *Human Capital*, New York 1964.  
 ——— and B. R. Chiswick, "Education and the Distribution of Earnings," *Amer. Econ. Rev. Proc.*, May 1966, 56, 358–69.  
 W. L. Hansen, "Total and Private Rate of Return to Investment in Schooling," *J. Polit. Econ.*, Aug. 1963, 71, 128–40.  
 ———, B. A. Weisbrod, and W. J. Scanlon, "Schooling and Earnings of Low Achievers," *Amer. Econ. Rev.*, June 1970, 60, 409–18.

<sup>6</sup> This effect is biased downward due to the truncation of the sample at  $AFQT=30$ .

<sup>7</sup> For an analysis of the option value of schooling, see Weisbrod.

- J. Mincer, "Schooling, Age and Earnings," *Human Capital and Personal Income Distribution*, Nat. Bur. Econ. Res., in progress 1972.
- N. A. Tolles, A. Hanson, and E. Clague, "The Structure of Economists Employment and Salaries, 1964," *Amer. Econ. Rev.*, Dec. 1965, supp., 55, 1-98.
- N. A. Tolles and E. Melichan, "Studies of the Structure of Economists Salaries and Income," *Amer. Econ. Rev.*, Dec. 1968, supp., 58, 1-154.
- B. Weisbrod, "Education and Investment in Human Capital," *J. Polit. Econ.*, Oct. 1962, supp., 70, 106-23.

# Schooling and Earnings of Low Achievers: Comment

By STANLEY MASTERS AND THOMAS RIBICH\*

In their recent article in this *Review*, W. Lee Hansen, Burton Weisbrod, and William Scanlon (HWS) are primarily concerned with estimating "the extent to which schooling, as contrasted with 'learning' or job training, is an important determinant of earnings of low achievers." In addition, they consider the net lifetime payoff to education (learning plus schooling per se) for low achievers and emphasize their conclusion that, at reasonable discount rates, the payoff is less than the cost of the education and the payoff rate less than the payoff rate to training.<sup>1</sup> Their analysis further suggests that the rate of return to education for low achievers is less than it is for the average student.<sup>2</sup> It is our contention that the calculations performed by HWS, with the data they use, do not provide convincing support for these latter inferences.

The HWS conclusions are based on a statistical analysis of an interview questionnaire given to 2,500 men, ages 17 to 25, who are low achievers, as measured by their scores on the Armed Forces Qualification Test (AFQT). In calculating the payoff to education, HWS take into account the effect of schooling on learning,<sup>3</sup> and the earnings

effect of that learning, as well as the earnings effect of schooling that is independent of the increase in learning. This measure of the payoff to education, combining the schooling and learning effects, is therefore practically the same as the estimates in the standard studies of this problem which do not make a distinction between the two effects. The only major difference is that HWS are focusing on low achievers, defined strictly as individuals who scored below the 30th percentile on the AFQT exam. The problems created by that special stratification seem to us to be more severe than HWS are willing to admit.

The qualification is made by HWS that their results for 17–25 year-old low achievers may very well understate the effect extra education produces for school-age low achievers, because "some youthful low achievers might have escaped this status by age 17–25, and did so as a result of subsequent schooling." Nevertheless, HWS argue that calculations with their sample at least answer directly the question "What has been the effect on income of differential amounts of schooling for men aged 17–25, who, following completion of their schooling, were judged to be low achievers?"<sup>4</sup> The HWS approach implies that, if we want to determine the economic value of (say) finishing high school for this group of young adult low achievers, we should look at the difference between the average earnings of high school graduates who are in the low-achiever sample and the average earnings of high school dropouts in the low-achiever sample. In our view, this type of comparison leads to an understatement of returns for the young adult low achievers (as well as for "school-age low achievers") and the understatement could be quite serious.

<sup>4</sup> It would seem more valuable to have accurate estimates on school-age low achievers since this is when policy action is normally undertaken. It might be possible, however, to predict ahead of time who is likely to be a young adult low achiever.

\* Visiting associate professor, University of Notre Dame and Institute for Research on Poverty, University of Wisconsin; and associate professor of economics, University of North Carolina at Chapel Hill, respectively. A National Science Foundation grant facilitated Ribich's work on this paper. We would like to thank Robert Gallman, Irwin Garfinkel, W. Lee Hansen, and Burton A. Weisbrod for helpful comments. Any remaining errors are our responsibility alone.

<sup>1</sup> Although HWS do not have comparable data on training costs, the returns to training are sufficiently large that the conclusion on comparable payoff rates would seem warranted.

<sup>2</sup> See Gary Becker, W. L. Hansen, Giora Hanoch, and Fred Hines et al. for studies that indicate a significantly higher rate of return for individuals aggregated by geography, race, and various education levels. HWS make no explicit statements in their conclusion about comparative rates of return between low achievers and others, but their interest in such a comparison is stated as on opening theme of their article.

<sup>3</sup> AFQT scores are used as a proxy for learning.

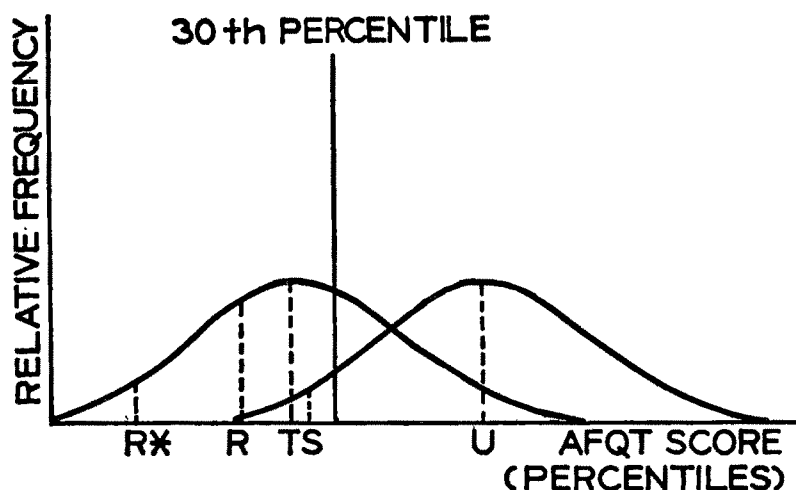


FIGURE 1

In Figure 1 the normal curve to the left represents an assumed frequency distribution of *AFQT* scores for *all* high school dropouts taking the *AFQT* test, and the normal curve to the right represents a similar frequency distribution of scores for all high school graduates. All those observations to the left of the heavy vertical line are included in the low-achiever subsample. The average test score for high school graduates who are low achievers is  $S$ , and the average score for low-achieving dropouts is  $R$ . Therefore, the distance  $RS$  represents the HWS estimate of the effects on learning of the extra schooling. The effects of increased learning on earnings can be estimated with the same data (as done by HWS) and added to the effect of schooling on earnings that is independent of changes in *AFQT* scores.

Note that, if the higher overall *AFQT* scores of graduates and dropouts are entirely the result of their extra schooling, as is assumed by the HWS calculations, then the frequency curve for the full sample of high school graduates would have coincided with the curve for the dropouts had the high school graduates not finished high school. The average extra learning acquired by all high school graduates is therefore indicated by the distance  $TU$ . It would seem fair to assume that a high school graduate with score  $S$  would have ended up in roughly the

same relative position on the curve if the graduates had not graduated and the graduate and dropout curve had coincided. In other words, we would expect him to have a score of  $R^*$ , which is less than  $R$ , the average score of the low-achieving dropout. His actual gain in learning as a result of high school is therefore better measured by  $R^*S$ , which is equal to  $TU$ .

Figure 2 puts the argument in the more appropriate terms of a linear regression. The line  $AB$  represents the regression relationship between years of schooling and *AFQT* scores for *all* individuals taking the *AFQT* test, and the line  $CD$  is the same relationship for the low-achiever subsample, which is the type of calculation provided by HWS. The regression line for the low-achiever subsample will not only lie below the regression line for the entire group (because the high *AFQT* scores are all eliminated when dealing with the low-achiever subsample) it will also have a smaller slope since the percentage of observations excluded from the low-achiever sample is larger, the greater the years of school.<sup>5</sup> The normal curves of Figure 1 are superimposed on Figure 2, with the means of these distributions assumed to be directly above the regression line  $AB$ . The

<sup>5</sup> We are assuming that the error term is (at least approximately) independent of the years of school.

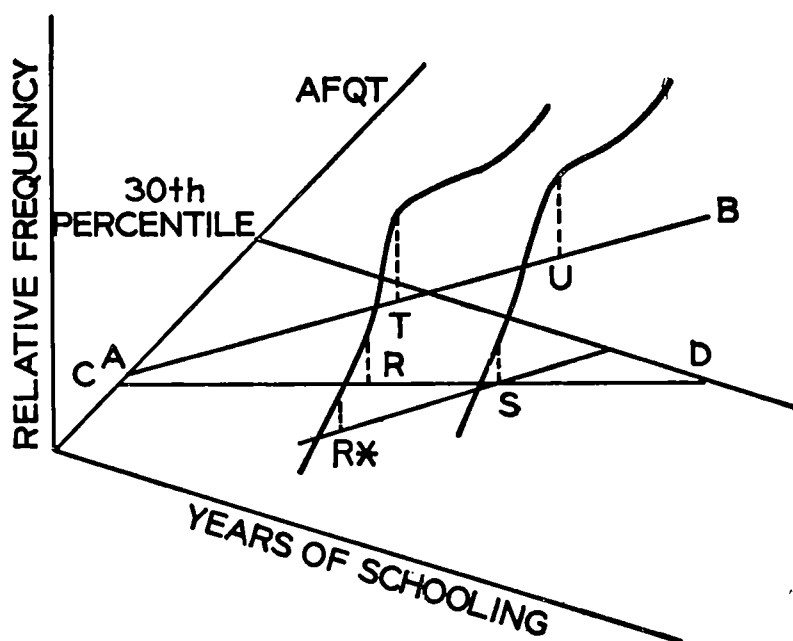


FIGURE 2

measured movement from  $R$  to  $S$  is now represented more generally by the  $CD$  regression line, while the movement from  $R^*$  to  $S$  turns out to be parallel with the regression line for the whole sample.

The interpretation of all this is clear enough. The relationship between years of schooling and learning for low achievers may, in reality, be very much the same as it is for others, despite the results of the HWS regressions. If the true relationship between schooling and learning is the same for low achievers as for others, if there is also a linear relationship between learning and earnings, and if the independent effect of years of schooling (holding learning constant) is the same for all groups, then the return to education for low achievers would come out exactly the same as it does for others. All these conditions may not be satisfied exactly, but there is nothing in the HWS calculations that seriously undermines them. The estimate of financial returns to low achievers is clearly biased downwards, and the HWS calculations do not preclude the possibility that returns to low achievers

are really about the same as they are for other groups.<sup>6</sup>

The rate of return for extra years of schooling, in other studies, comes out to be (most frequently) about 15 percent for schooling levels below college. Remarkably, those studies have indicated fairly small rate-of-return differences, at these schooling levels, among individuals from different regions of the country and of different races; yet average achievement levels for the South and non-South and for whites and non-whites are markedly different. The relationship between average achievement levels for various groups and the returns experienced from extra years of schooling is generally

<sup>6</sup> Other studies could easily fall into an error similar to that of HWS since the HWS data are not the only survey information concentrating on some group of the population that has not "made it." For example, the *Survey of Economic Opportunity*, which includes a disproportionate sampling from poor neighborhoods, is now being widely used for diverse purposes. If one were to apply standard regression techniques to calculate the payoff to education for the *Survey's* large sample of individuals living in poor neighborhoods, this procedure would lead to an underestimate similar to that of HWS.

unimpressive.<sup>7</sup> This suggests that the exceedingly low returns calculated by HWS may be due chiefly to the peculiar stratification of their sample, which excludes all who have learned more than some arbitrary amount.<sup>8</sup> The rate of return to extra years of schooling for low achievers, reasonably defined and measured in the customary way, may be much closer to the 15 percent average than it is to the rate of return of well under 5 percent calculated by HWS.<sup>9</sup>

Important policy decisions might very well hinge on this issue. If, for instance, we are faced with the problem of allocating some given amount of extra educational resources among students of varying ability, the above analysis suggests (contrary to the HWS results) that the argument of "economic ef-

ficiency" cannot be used to justify disproportionate allocations to those of high ability. Furthermore, unless very large investments are being considered such that appreciable diminishing returns set in, concentrating new educational inputs among those who are faltering in school cannot be dismissed on efficiency grounds. Such concentration would surely be desirable on the grounds of equal educational opportunity and an improved income distribution, since those with low measured learning are most often those disadvantaged by the low socioeconomic circumstances of their homes and neighborhoods; hence neutrality in terms of economic efficiency might indicate that new budget allocations should indeed be concentrated among low achievers.

Note, however, that even though the rate of return to education for low achievers may be as high as for other groups and greater than any plausible discount rate, it is still quite possible that the costs exceed the economic benefits for available social policies aimed at increasing the amount of general education. First, augmenting the flow of students through additional years of schooling will usually involve either some supplementary costs (for counseling and the like) or lower learning gains for these marginal students. Weisbrod's well-known study of a model dropout-prevention program indicates that, once we take account of the extra costs required to reduce the dropout rate, the total cost of rescuing and educating a potential dropout may be well above the anticipated earnings gain.<sup>10</sup> Second, the portion of the HWS analysis which separates learning effects from the sheepskin effect of years of schooling suggests strongly that simply forcing or bribing individuals through extra schooling, without assurance that learning takes place, will fail to have a satisfactory payoff rate.<sup>11</sup> Finally, other

<sup>7</sup> See Hanoch, Table 3, and Hines et al., Tables 2 and 4. In particular, note the following interesting results. Hines et al., find that, among males who have completed eight years of school, additional education through high school yields a *higher* rate of return for nonwhites than for whites. In addition, they find that, for education below college, the social rate of return is greater in the South than in other regions. Similar results are obtained by Hanoch. For those outside the South with eight years of school, he also finds that additional education through high school yields a higher rate of return for nonwhite than for white males. And at these schooling levels, he finds that white males earn a higher rate of return on additional education in the South than elsewhere. Therefore in these cases (which are among the most relevant for the HWS sample), the returns to education are higher for groups with lower scores on tests like the *AFQT*.

<sup>8</sup> There is also another reason why the returns calculated by HWS are relatively low. They assume that the earnings differential observed at ages 17-25 among individuals with different levels of education will simply stay the same through the rest of the individuals' working lives. This assumption is clearly out of accord with the consistently observed tendency for such differentials to grow appreciably throughout life. Other studies (see fn. 7) commonly use cross-section data, including individuals whose earnings are at a career peak, and often take into account secular growth in differentials due to increasing productivity. It follows that HWS' neglect of such adjustments is an added reason why their estimates of payoffs to education are lower than the results of most other studies.

<sup>9</sup> HWS do not calculate internal rates of return but rather discount returns by 5 and 10 percent. The 5 percent discount rate yields a present value of only one-half the lower-bound estimate of the costs of additional schooling, thereby suggesting an internal rate of return of less than 3 percent.

<sup>10</sup> See Weisbrod. In this study the earnings differentials between individuals with different levels of education, as reported in the Census, are used as the basis for the returns calculation. It is those same differentials that yield quite high rates of return when the extra costs of dropout prevention programs are not included in the estimate.

<sup>11</sup> A major conclusion of HWS, that schooling per se

calculations indicate that simply spending more money on compensatory education or on additional amounts of standard school inputs also produces returns that are below costs.<sup>12</sup> Therefore the HWS conclusions that governmental efforts in the realm of general education have a lower financial payoff rate than training programs and yield financial returns that (when discounted) are less than costs are both probably correct, though for more complex reasons than presented in their article.

has less effect on earnings than does the learning associated with that schooling, is strengthened by taking account of the downward bias in their estimate of the effect of schooling on learning. This strengthening might actually be needed to fully preserve this conclusion, since schooling per se has a larger earnings effect than the associated learning in the HWS regression model where the training variable is excluded—a model which may be more appropriate for considering the relative effects of learning and schooling per se since a person's chances of selecting and/or being selected for training can be legitimately thought of as part of the benefit of either learning or schooling.

<sup>12</sup> See Ribich, ch. 4 and Appendix C.

# REFERENCES

- G. S. Becker, *Human Capital*, New York 1964.
- G. Hanoch, "An Economic Analysis of Earnings and Schooling," *J. Hum. Resources*, summer 1967, 2, 310-19.
- W. L. Hansen, "Total and Private Rates of Return to Investment in Schooling," *J. Polit. Econ.*, Apr. 1963, 71, 128-40.
- , B. A. Weisbrod, and W. J. Scanlon, "Schooling and Earnings of Low Achievers," *Amer. Econ. Rev.*, June 1970, 60, 409-18.
- F. Hines, L. Tweeten, and M. Redfern, "Social and Private Rates of Return to Investment in Schooling, By Race-Sex Groups and Regions," *J. Hum. Resources*, summer 1970, 5, 318-40.
- T. Ribich, *Education and Poverty*, Washington 1968.
- B. A. Weisbrod, "Preventing High School Dropouts," in R. Dorfman, ed. *Measuring Benefits of Government*, Washington 1965, 117-49.

# Schooling and Earnings of Low Achievers: Reply

By W. LEE HANSEN, BURTON A. WEISBROD, AND WILLIAM J. SCANLON\*

Our recent paper in this *Review* is no exception to the rule that every empirical study can be expected to generate two types of critical comments: 1) that the data are less than satisfactory, and 2) that the model is not well specified.

Stanley Masters and Thomas Ribich (M-R) tackle us primarily on the first ground, claiming that our stratification by Armed Forces Qualification Test (AFQT) score poses problems that are "more severe than [we] are willing to admit." Let us consider their argument. To begin with, M-R are wrong in asserting that our analysis assumed that "... the higher overall AFQT scores of graduates and dropouts are entirely the result of their extra schooling. . . ." We made no such assumption; to the contrary, we stated (albeit in fn. 13) that our model of the determinants of AFQT scores included, in addition to schooling, "the independent variables in Model IV . . .", namely training, age, color, marital status, family size, and region of residence. Consequently, the conclusion by M-R that "the frequency curve for the full sample of high school graduates would have coincided with the curve for the dropouts had the high school graduates not finished high school" does not follow.

Even if we grant this M-R argument, however, we question their statement that "It would seem fair to assume that a high school graduate with score *S* would have ended up in roughly the same relative position on the dropout curve if the graduates

had not graduated. . . ." This is a crucial assumption, but one that is by no means self-evident. Given the selection process that distinguishes persons attaining various levels of schooling, there is little reason to believe that a person who changed his level of schooling would retain his relative income position among persons with the new educational level. Consequently, we cannot agree with M-R that our estimate of financial returns to low achievers is "clearly biased downwards." Nevertheless they are quite right in stating that our empirical estimates "*do not preclude* the possibility that returns to low achievers are really about the same as they are for other groups" (emphasis added). Perhaps a statement such as theirs should appear at the end of every empirical paper that compares responses of population groups—namely that any findings of differences in responses "*do not preclude*" the possibility that the responses are, truly, the same. But, then, neither do findings of *no* differences preclude the possibility that the responses are, truly, different.

Our objective was to go beyond the studies of financial returns from schooling for large population aggregates in order to learn more about the financial benefits from schooling for particular subsets of the population—in this case, the subset of "low achievers." Our presumption is that every subset does not realize the mean return—and surely M-R would not dispute this as a general proposition. Yet their assertion that returns to low achievers "... may be much closer to the 15 percent average than it is to the rate of return of well under 5 percent calculated by HWS. . ." might just as well be made for each and every other group for which below-average (or above-average) returns might be found. Conjecture—theirs or ours—is no substitute for research, however, and while our estimates do suggest to us that low achievers may not gain as much

\* Hansen and Weisbrod are professors of economics and of educational policy studies, and senior staff members, Institute for Research on Poverty, University of Wisconsin; Scanlon is research staff member at the Urban Institute. This is a portion of a larger study dealing with the relationships among education, ability, and income, supported by the Ford Foundation. We wish to acknowledge the financial support provided by the Ford Foundation and the Institute for Research on Poverty, University of Wisconsin.

from schooling as do other students, our study is hardly the final word.<sup>1</sup>

Barry Chiswick's comment amounts to a statement that we have misspecified the model of earnings determinants by including age rather than "labor market experience." "By holding age constant," he writes, "an additional year of schooling means a year less of experience."

The issue is an important one, although less simple than Chiswick implies. It is true that a person who attends school does not get the *same* labor market experience that he could if he were on the job, but this fact does not resolve the question of whether age or years-since-leaving-school (Chiswick's proposed measure of "experience") is a better proxy for the kinds of nonschooling experiences that contribute to earning power. Not only do persons in school often hold part-time and summer jobs, but they gain a wide variety of other experiences during the hours and days they are not attending to schoolwork which may be of subsequent value in the labor market. Age is, after all, a proxy for a host of experiences that may affect subsequent earnings; and years-since-leaving-school is a proxy for another set of experiences affecting earnings.

The point is that *both* age and experience-on-the-job probably affect earnings. The importance of the issue posed by Chiswick, however, is not whether age or experience is more important, but, rather, what the effect is on the coefficient of *schooling* when experience is substituted for age. As he points out, the elimination of age in the regression equation, and the substitution of experience (age minus schooling minus 5) leads to an enormous increase in the coefficient of schooling.

Chiswick's proposed measure of experience is, however, only a first approximation to labor market experience and, hence, experience *on the job*, since it assumes implicitly that the individual is fully employed subsequent to leaving school. This assumption

TABLE 1

	Original Results (HWS Model IV)	Alternative Experience Measures		
		Years Since Leaving School (Chiswick Model)	Years Since Leaving School After Age 14	Years Since Leaving School After Age 16
Education	\$20 (1.54)	\$204	\$89 (6.08)	\$42 (4.13)
Age	184 (12.09)	—	—	—
Experience	—	184	161 (11.10)	180 (11.90)

*Note:* *t*-values shown in parentheses. All estimates are for Model IV in our earlier paper and, hence, control for the following variables: *AFQT*, training, color, marital status, family size, and region of residence. Only the direct effects of education are shown; the effect of education on "learning" (*AFQT*) and, thus, indirectly on earnings is not included in these estimates.

tion is certainly inaccurate for young male high-school dropouts. As of March 1965, less than 80 percent of this school dropout group (age 18–24) was in the labor force, and 14 percent of those were unemployed. (See "Labor Market Twist.") Thus, fewer than 69 percent of these men were employed at the time, obtaining work experience. The assumption, implicit in Chiswick's measure of "labor market experience," that each person is fully employed at all times subsequent to leaving school is likely to be even more inaccurate for those dropouts who are at the bottom of the achievement ladder, as are all the men in our sample. The reluctance of employers to hire high school dropouts, along with the impact of child labor laws, particularly restrict the work experience of young people prior to age 16. Thus, to determine the amount of time during which a person was obtaining experience *on the job* we would, ideally, need to know the proportion of "full-time" that each man in our sample worked since leaving school.<sup>2</sup> Unfortunately, no such information is available.

We have constructed an experience measure which takes some account of the difference between obtaining work experience

<sup>1</sup> In an earlier paper (see Hansen, Allen Kelley, and Weisbrod) we discussed and provided some evidence on the differential effects of schooling on learning-achievement for various students.

<sup>2</sup> In saying this we are assuming that any labor market experience obtained by a person who is either unemployed or not in the labor force does not affect his productivity and subsequent earnings.

and simply being out of school. Our measure assumes that no "labor market experience" is acquired prior to age 16, or alternatively, age 14. The resulting estimated coefficients for education and for these measures of experience—with experience defined as years, subsequent to age 14 (16) and subsequent to leaving school—appear in columns 3 and 4 of Table 1. Our original results (shown in column 1) assume that, at one extreme, age and years of schooling have separate and independent effects on earnings. The results based on Chiswick's definition of experience (shown in column 2) assume, by contrast, that age has no effect independent of years of schooling or experience. The results based on our alternate views of experience assume intermediate positions, that no work experience is obtained before age 14 (column 3) or 16 (column 4) even if the individual is out of school.

As Chiswick states, his approach provides an "upper limit estimate of the effect of schooling"—\$204 versus our original estimate of \$20 of direct effect. At the same time, his experience coefficient takes on the value of \$184 which was previously attributed to age. Use of our modified experience variable, however, leads to education coefficients substantially reduced from Chiswick's estimate: \$89 per year when experience is permitted to begin at age 14, and \$42 when experience is permitted to begin at age 16. These estimates are well below Chiswick's figure of \$204, and far closer to our figure of \$20.

If our experience measures can be viewed as more reasonable proxies for labor market experience than Chiswick's, which counts as experience all years not spent in school, then the rate of return to additional schooling is still low, though not quite as low as our earlier results may have suggested. Using

Chiswick's approach to estimating a rate of return, our Model IV estimates yield a return of approximately 6 percent when experience begins at age 14 and 2 percent when experience begins at age 16; these rates of return are far below Chiswick's 12 percent but at the same time above the 1 percent he estimated from our original results.

We hope that further work will be undertaken on the importance of experience, but as it is, careful attention should be paid to defining that term.<sup>3</sup> For some jobs, age rather than experience "on the job" may be of critical importance; in other jobs the actual amount of time spent on that type of work will be of key importance. Ideally, independent measures of both age and experience would be obtained and introduced into models of the determinants of earnings.

#### REFERENCES

- B. Chiswick, "Schooling and Earnings of Low Achievers: Comment," *Amer. Econ. Rev.*, 1972, 62, 752-54.
- W. L. Hansen, B. A. Weisbrod, and W. J. Scanlon, "Schooling and Earnings of Low Achievers," *Amer. Econ. Rev.*, June 1970, 60, 409-18.
- W. L. Hansen, A. C. Kelley, and B. A. Weisbrod, "Economic Efficiency and the Distribution of Benefits from College Instruction," *Amer. Econ. Rev. Proc.*, May 1970, 60, 335-40.
- S. Masters and T. Ribich, "Schooling and Earnings of Low Achievers: Comment," *Amer. Econ. Rev.*, 1972, 62, 755-59.
- U.S. Department of Labor, "The Labor Market 'Twist,' 1964-1969," Special Labor Force Report 133, 1971, p. A-3.

<sup>3</sup> We examine alternative experience measures further in a paper now nearing completion.

## Statement of Editorial Policy

Anyone who has read the *American Economic Review* for the last twenty-five years realizes that great changes have taken place in the subject matter and methodology of economic research. Articles on mathematical economics and the finer points of economic theory occupy a much more prominent place than ever before, while articles of a more empirical, policy-oriented, or problem-solving character seem to appear less frequently. Such changes have occurred slowly, and are to a degree illusory. Nevertheless, readers may wonder and some have inquired whether the editor of the *Review* and his colleagues, the Board of Editors, have sought to further this trend, or whether we have a prior preference for one type of manuscript over another. No such preference exists. The pages of the journal are open to every subdivision of economics and to every method. We are governed in what we publish first by the papers that are submitted and second by those that survive the editorial refereeing. Space allows us to print only 14 percent of the papers submitted. Thus, we seek to make editorial refereeing as objective as possible. An acceptable paper must either advance the art of economic analysis, or else exemplify a highly skillful and pertinent application of that art. The function of the *Review* is to publish those papers which contain the most fruitful research, the most lucid discussions, and which open the most promising avenues for further inquiry.

While much published work contains mathematical models, this must not hide the fact that many of these papers are directed to applied subjects such as government stabilization and taxation, monetary theory, financial institutions, international trade, public regulation, racial discrimination, poverty, health, education, housing, environmental control, and economic development. The *Review* welcomes theoretical and empirical and policy-oriented papers in all areas of economic inquiry, and recognizes the validity of a variety of methods of research and exposition. The distribution of space in the journal is highly elastic and reflects the quality and volume of work in various fields, not any prior judgments as to their importance.

When the American Economic Association was founded in 1885, its report of organization stated three major objectives: "The encouragement of economic research, the publication of economic monographs, and the encouragement of perfect freedom in all economic discussion." These goals retain their vitality and the present statement of editorial policy reaffirms them.

GEORGE H. BORTS  
Managing Editor

# NOTES

## EIGHTY-FIFTH ANNUAL MEETING OF THE AMERICAN ECONOMIC ASSOCIATION

Toronto, Ontario, Canada, December 27–30, 1972

### *Preliminary Announcement of the Program*

Wednesday, December 27, 1972

2:00 P.M. EXECUTIVE COMMITTEE MEETING

Thursday, December 28, 1972

8:30 A.M. RADICAL ECONOMICS AND THE HISTORY OF ECONOMIC THOUGHT

*Chairman* WARREN J. SAMUELS, Michigan State University

*Papers:* DAN FUSFELD, University of Michigan

Types of Radicalism in American Economics

EDWARD J. NELL, New School for Social Research

Chapter 25 of *Capital* and the Radical Critique of Contemporary Growth Theory

*Discussants:* MARTIN BRONFENBRENNER, Duke University

ROBERT EAGLY, University of North Carolina, Chapel Hill

E. K. HUNT AND HOWARD J. SHERMAN, University of California, Riverside

8:30 A.M. SPATIAL ECONOMICS

*Chairman:* Mahlon R. Straszheim, University of Maryland

*Papers:* CHARLES WOLF, JR. AND DAVID WEINSCHROTT, Rand Corporation

International Transactions, "Regionalism": Criteria for Distinguishing "Insiders" from "Outsiders"

EYTAN SHESHINSKI, Hebrew University

Optimal Population Distribution in the City

*Discussants:* GEORGE TOLLEY, University of Chicago

PAUL WONNACOTT, University of Maryland

JEROME ROTHENBERG, Massachusetts Institute of Technology

8:30 A.M. ECONOMETRIC MODELS

*Chairman:* T. Merritt Brown, University of Western Ontario

*Papers:* LAWRENCE R. KLEIN, University of Pennsylvania AND GARY FROMM, The Brookings Institution

A Comparison of Nine Econometric Models of the United States

HIROKI TSURUMI, Queen's University

A Comparison of Econometric Macro Models of the United States, Canada, and Japan

CHIKASHI MORIGUCHI, Osaka University

Project LINK

*Discussants:* RONALD G. BODKIN, Economic Council of Canada

GEORGE R. SCHINK, Wharton Econometric Forecasting Associates, Inc.

KANTA MARWAH, Carleton University

8:30 A.M. RACIAL ASPECTS OF INDUSTRIAL EMPLOYMENT\* (Joint Session with Association for the Study of the Grants Economy)

*Chairman:* GEORGE M. VON FURSTENBERG, Indiana University

*Papers:* DURAN BELL, The Brookings Institution

The Racial Structure of Industrial Employment

ORLEY ASHENFELTER, Princeton University and JAMES HECKMAN, Columbia University and the National Bureau of Economic Research

Estimating the Effect of the Federal Government on Racial Discrimination in Labor Markets

*Discussants:* BRADLEY R. SCHILLER, University of Maryland

A. MICHAEL SPENCE, Harvard University

10:30 A.M. MICROPOLITICS AND MACROECONOMICS (Joint Session with the Public Choice Society)

*Chairman:* GORDON TULLOCK, Virginia Polytechnic Institute

\* Not to appear in the *Papers and Proceedings*

*Papers:* GEORGE J. STIGLER, University of Chicago

Empirical Work on Relations Between Voting and General Economic Conditions

PAUL McCracken, University of Michigan

The Practice of Political Economy

ARTHUR OKUN, The Brookings Institution

Comment on Stigler's Paper

*Discussants:* WILLIAM RIKER, University of Rochester

THOMAS IRELAND, University of Virginia

#### 10:30 A.M. THE EXPERIENCE OF DEVELOPMENT: LESSONS FOR THEORY AND POLICY

*Chairman:* IRMA ADELMAN, University of Maryland

*Papers:* JEFFREY WILLIAMSON, University of Wisconsin and ALLEN C. KELLEY, Duke University

Dualistic Theories and Quantitative Analysis: The Contemporary Relevance of Meiji Japanese History

HOLLIS B. CHENERY AND N. G. CARTER, International Bank for Reconstruction and Development  
An Evaluation of Development Performance, 1960-70

*Discussants:* DONALD J. HARRIS, University of Wisconsin

GUSTAV RANIS, Yale University

#### 10:30 A.M. ECONOMICS OF POLLUTION

*Chairman:* ALLEN V. KNEESE, Resources for the Future, Inc.

*Papers:* CLIFFORD RUSSELL, Resources for the Future, Inc.

Application of a Micro-Economic Model to Regional Environmental Management

WALTER ISARD, University of Pennsylvania

Application of Input-Output and other Regional Science Models to Environmental Management

MICHAEL K. EVANS, Chase Econometric Associates, Inc.

Application of an Econometric Forecasting Model to Pollution Control Costs

*Discussants:* ROBERT DORFMAN, Harvard University

#### 12:30 P.M. PLENARY LUNCHEON\* (Allied Social Science Associations)

*Chairman:* (To be announced)

*Speaker:* A Canadian Cabinet Minister

#### 2:30 P.M. RACIAL AND OTHER DISCRIMINATION

*Chairman:* BERNARD ANDERSON, University of Pennsylvania

*Papers:* FINIS WELCH, City University of New York and National Bureau of Economic Research

The 1960 and 1970 Census Results

JOSEPH E. STIGLITZ, Yale University

Conceptual Approaches to the Economics of Discrimination

MARCUS ALEXIS, Northwestern University

A Theory of Labor Market Discrimination With Interdependent Utilities

*Discussants:* THOMAS SOWELL, University of California, Los Angeles

BARBARA BERGMANN, University of Maryland

#### 2:30 P.M. NATURAL RESOURCES AS A CONSTRAINT ON ECONOMIC GROWTH

*Chairman:* ANTHONY D. SCOTT, University of British Columbia

*Papers:* H. SCOTT GORDON, University of Indiana and Queen's University

Today's Apocalypses and Yesterday's

NATHAN ROSENBERG, University of Wisconsin

Innovative Responses to Materials Shortages

PAUL BRADLEY, University of British Columbia

Increasing Scarcity: The Case of Energy Resources

*Discussants:* MASON GAFFNEY, University of Wisconsin-Milwaukee and Resources for the Future, Inc.

CHANDLER MORSE, Cornell University

#### 2:30 P.M. THE CHINESE ECONOMY AT THE PRESENT JUNCTURE (Joint Session with the Association for Comparative Economic Study)

*Chairman:* GREGORY GROSSMAN, University of California, Berkeley

*Papers:* TA-CHUNG LIU, Cornell University and K. C. YEH, Rand Corporation

Chinese and Other Asian Economies: A Quantitative Evaluation

DWIGHT PERKINS, Harvard University

Economic Organization and Policy of China

\* Not to appear in the *Papers and Proceedings*

*Discussants:* MORRIS BORNSTEIN, University of Michigan  
 S. H. CHOU, University of Pittsburgh  
 RICHARD S. ECKAUS, Massachusetts Institute of Technology  
 CARL RISKIN, Columbia University

2:30 P.M. THE MULTI-NATIONAL FIRM: BANE OR BOON?\* (Joint Session with the American Finance Association)

*Chairman:* SYLVIA OSTRY, Economic Council of Canada

*Papers:* RAYMOND VERNON, Harvard University

Perspectives on Foreign Direct Investment: Capital Exporting Country

ALBERT E. SAFARIAN, University of Toronto

Perspectives on Foreign Direct Investment: Capital Receiving Country

ANDREW F. BRIMMER, Board of Governors, Federal Reserve System

Multi-National Banks and the Management of Monetary Policy

*Discussants:* ANDREAS G. PAPANDREOU, York University

C. FRED BERGSTEN, The Brookings Institution

GUY E. NOYES, Morgan Guaranty Trust Co.

2:30 P.M. ISSUES IN GRANTS ECONOMICS\* (Joint Session with the Association for the study of the Grants Economy)

*Chairman:* KENNETH E. BOULDING, University of Colorado

*Papers:* GERALD I. WEBER, University of California, Berkeley

Grant Elements in the Financing of Elementary and Secondary Education

ROBERT C. BUSHNELL, Wayne State University

Negative Grants and Welfare Losses Due to Copelessness among the Aged

THOMAS C. IRELAND, Loyola University

Transfers in Kind: The Case of the Blood Bowl

*Discussants:* JANOS HORVATH, Butler University

HENRY AARON, The Brookings Institution

JEREMY WARFORD, The International Monetary Fund

8:00 P.M. RICHARD T. ELY LECTURE (Joint Session with the Econometric Society)

*Chairman:* KENNETH J. ARROW, Harvard University

*Speaker:* LEONID HURWICZ, University of Minnesota

The Design of Mechanisms for Resource Allocation

Friday, December 29, 1972

8:30 A.M. INTERGENERATIONAL DETERMINANTS OF INDIVIDUAL INCOMES

*Chairman:* SHIRLEY JOHNSON, Vassar College

*Papers:* JOHN BRITTAIN, The Brookings Institution

The Transmission of Material Wealth

SAMUEL BOWLES, Harvard University

Effects of IQ, Schooling, and Social Class

*Discussants:* MARTIN DAVID, University of Wisconsin

JACOB MINCER, Columbia University and National Bureau of Economic Research

JAMES SMITH, Pennsylvania State University

8:30 A.M. ECONOMICS OF INFORMATION

*Chairman:* GEORGE J. STIGLER, University of Chicago

*Papers:* JACK HIRSHLEIFER, University of California, Los Angeles

Where are We in the Theory of Information?

LESTER TELSER, University of Chicago

Searching for the Lowest Price

*Discussants:* PHILLIP NELSON, State University of New York, Binghamton

MICHAEL ROTHSCHILD, Princeton University

8:30 A.M. ECONOMICS OF HEROIN

*Chairman:* RICHARD ZECKHAUSER, Harvard University

*Papers:* RAUL FERNANDEZ, University of California, Irvine

The Problem of Heroin Addiction and Radical Political Economy

CHRISTOPHER CLAGUE, University of Maryland

The Operation of the Heroin Market Under Alternative Systems of Control

MARK MOORE, Harvard University

Law Enforcement to Achieve Discrimination on the Effective Price of Heroin

\* Not to appear in the *Papers and Proceedings*

*Discussants:* JAMES SEAGRAVES, University of North Carolina, Raleigh-Durham  
 GARY BECKER, University of Chicago (tentative)  
 LESTER LAVE, Carnegie-Mellon University

8:30 A.M. Trade Liberalization in Agricultural Products\* (Joint Session with the American Agricultural Economics Association)

*Chairman:* T. K. Warley, University of Georgia

*Papers:* FREDERICK BERGSTEN, The Brookings Institution

Future Directions for United States Trade

JOHN A. SCHNITTKER, Schnittker Associates, Washington, D.C.

Prospects for Freer Agricultural Trade

D. GALE JOHNSON, University of Chicago

The Impact of Freer Trade on North American Agriculture

*Discussants:* DALE HATHAWAY, Michigan State University

PIERRE MALVÉ, Councillor for Trade Affairs, European Communities

G. I. TRANT, Director General, Canada Department of Agriculture

10:30 A.M. WOMEN ECONOMISTS: GUIDELINES FOR THEIR EDUCATION AND EMPLOYMENT\*

*Chairman:* CAROLYN SHAW BELL, Wellesley College

*Panel:* A.E.A. COMMITTEE ON THE STATUS OF WOMEN IN THE ECONOMICS PROFESSION:

WALTER ADAMS

COLLETTE MOSER

FRANCINE BLAU

BARBARA REAGAN

MARTHA BLAXALL

MYRA STROBER

KENNETH BOULDING

PHYLLIS WALLACE

JOHN KENNETH GALBRAITH

*Discussion:* Comments invited from the floor on proposed guidelines for action

10:30 A.M. DUAL LABOR MARKETS (Joint Session with the Industrial Relations Research Association)

*Chairman:* LEONARD A. RAPPING, Carnegie-Mellon University

*Papers:* MICHAEL REICH, Boston University, DAVID GORDON, Harvard University, AND RICHARD EDWARDS, Harvard University

A Radical Theory of Labor Market Segmentation

THOMAS VIETORISZ, New School for Social Research AND BENNETT HARRISON, Massachusetts Institute of Technology

A Theory of Subemployment and the Labor Market

MICHAEL PIRE, Massachusetts Institute of Technology

Labor Market Stratification and Wage Determination

*Discussants:* JAMES CROTTY, Bucknell University

FRANCINE BLAU, Trinity College

ROBERT ARONSON, Cornell University

10:30 A.M. BEHAVIOR AND ENTRY UNDER PUBLIC UTILITY REGULATION (Joint Session with the Transportation and Public Utilities Group)

*Chairman:* HARRY M. TREBING, Michigan State University

*Papers:* LELAND L. JOHNSON, Rand Corporation

Behavior of the Firm Under Regulatory Constraint: A Reassessment

WILLIAM G. SHEPHERD, University of Michigan

Entry as a Substitute for Regulation

*Discussants:* JAMES R. NELSON, Amherst College

DALLAS W. SMYTHE, University of Saskatchewan

ELIZABETH E. BAILEY, Bell Laboratories

10:30 A.M. MATHEMATICAL ECONOMICS IN THE SOVIET UNION\* (Joint Session with the Association for Comparative Economic Studies)

*Chairman:* VLADIMIR G. TREML, Duke University

*Papers:* RICHARD JUDY, University of Toronto

The Contribution of Computers and Mathematical Economics to Soviet Planning and Management

LEON SMOLINSKI, Boston College

The Formative Period of Soviet Mathematical Economics

*Discussants:* ROBERT W. CAMPBELL, Indiana University

HERBERT S. LEVINE, University of Pennsylvania

\* Not to appear in the *Papers and Proceedings*

10:30 A.M. THE COSTS AND BENEFITS OF THE DOLLAR AS A RESERVE CURRENCY\* (Joint Session with the American Finance Association)

*Chairman:* RICHARD A. LABARGE, Louisiana State University

*Papers:* PETER KENEN, Princeton University

Convertibility and Consolidation: A Survey of Options for Future Reform of the System

PAUL McCracken, University of Michigan

Convertibility and Domestic Economic Policy

JOHN HELLIWELL, University of British Columbia

Dollars as Reserve Assets: The Creditor's Viewpoint

*Discussants:* ROBERT M. MACINTOSH, Bank of Nova Scotia, Toronto

H. R. HELLER, University of Hawaii

12:30 P.M. JOINT LUNCHEON\* (with the American Finance Association)

*Chairman:* (To be announced)

*Speaker:* ARTHUR F. BURNS, Chairman, Board of Governors, Federal Reserve System

2:30 P.M. ECONOMIC EDUCATION

*Chairman:* G. LELAND BACH, Stanford University

*Papers:* A.E.A. COMMITTEE ON ECONOMIC EDUCATION

An Agenda for Improving Teaching in Economics

KENNETH BOULDING, University of Colorado

Introducing Freshmen to the Social System

JOHN J. SIEGFRIED, Vanderbilt University and KENNETH WHITE, Rice University

Financial Rewards to Research and Teaching: A Case Study of Academic Economists

*Discussants:* R. A. GORDON, University of California, Berkeley

LESTER FETTER, Ocean County College

NANCY GORDON, Carnegie-Mellon University

2:30 P.M. SPECIAL INVITED LECTURE

*Chairman:* HARRY G. JOHNSON, University of Chicago

*Speaker:* RICHARD KAHN, Cambridge University

The International Monetary System

8:00 P.M. PRESIDENTIAL ADDRESS

*Chairman:* WALTER GORDON, Former Minister of Finance, Canada

*Speaker:* J. KENNETH GALBRAITH, Harvard University

9:15 P.M. BUSINESS MEETING

Saturday, December 30, 1972

8:30 A.M. ECONOMICS OF POPULATION

*Chairman:* MARC NERLOVE, University of Chicago

*Papers:* T. PAUL SCHULTZ, University of Minnesota

Economic Factors Affecting Population Growth

JOSEPH L. FISHER and RONALD RIDKER (tentative), Resources for the Future, Inc.

Effects of Population Growth on Resource Availability and Environmental Quality

*Discussants:* D. GALE JOHNSON, University of Chicago

NORMAN B. RYDER, Princeton University

KENNETH BOULDING, University of Colorado

8:30 A.M. DECISION MAKING UNDER UNCERTAINTY

*Chairman:* DAVID STUHR, Columbia University

*Papers:* AARON J. DOUGLAS, Harvard University

Stochastic Returns and the Theory of the Firm

STEPHEN A. ROSS, University of Pennsylvania

Economic Theory of Agency

URI BEN-ZION, University of Minnesota and JAMES L. BICKSLER, Rutgers University

Portfolio Choice with Market Information: A Model of an Active Investor

*Discussants:* HAYNE LELAND, Stanford University

WILLIAM A. BROCK, University of Rochester

ROBERT GLAUBER, Harvard University

\* Not to appear in the *Papers and Proceedings*

- 8:30 A.M. COMPARATIVE VIEW OF DISTRIBUTION OF INCOME AND WEALTH\* (Joint Session with the Association for Comparative Economic Studies)  
*Chairman:* JANET G. CHAPMAN, University of Pittsburgh  
*Papers:* ANDRZEJ BRZESKI, University of California, Davis  
 Income Distribution under Socialism: Poland  
 JAN M. MICHAL, State University of New York, Binghamton  
 Size Distribution of Earnings and Personal Income in Czechoslovakia, Hungary, and Yugoslavia  
 MASATASHI KURATANI, Japanese Ministry of Finance and University of Chicago  
 Income Distribution in Japan  
*Discussants:* JACOB MINCER, Columbia University and the National Bureau of Economic Research  
 MARTIN BRONFENBRENNER, Duke University
- 8:30 A.M. THE DYNAMICS OF THE URBAN PROPERTY MARKET\* (Joint Session with the American Real Estate Urban Economics Association)  
*Chairman:* HUGH O. NOURSE, University of Missouri, St. Louis  
*Papers:* ROBERT BURCHELL, JAMES HUGHES, AND FRANKLIN JAMES, Rutgers University  
 Urban Change, Neighborhood Deterioration, and Housing Abandonment  
 BRUCE HAMILTON, Princeton University  
 Zoning and Property Taxation in a System of Governments  
 GREG INGRAM AND JOHN KAIN, Harvard University  
 Representing Housing Disinvestment and Abandonment with the NBER Urban Simulation Model  
*Discussants:* FRANK DELEEUW, The Urban Institute  
 JOSEPH S. DESALVO, University of Wisconsin, Milwaukee
- 10:30 A.M. ORGANIZATIONAL FORMS AND INTERNAL EFFICIENCY  
*Chairman:* ROLAND N. MCKEAN, University of Virginia  
*Papers:* OLIVER E. WILLIAMSON, University of Pennsylvania  
 Primary Work Group and Transactional Theories of the Firm  
 DONALD L. MARTIN, University of Virginia  
 Other Organizational Forms and their Implications  
*Discussants:* KAREN DAVIS, The Brookings Institution  
 EDWIN G. DOLAN, Dartmouth College  
 JOSEPH D. REID, University of Pennsylvania
- 10:30 A.M. INTERPRETATIONS OF ECONOMIC GROWTH  
*Chairman:* LIONEL MCKENZIE, University of Rochester  
*Papers:* MOSES ABRAMOVITZ AND PAUL DAVID, Stanford University  
 An Interpretation of Nineteenth-Century Growth in the United States  
 RICHARD R. NELSON, Yale University AND SIDNEY G. WINTER, JR., University of Michigan  
 Economic Growth as an Evolutionary Process  
*Discussants:* HAZEL DENTON, Harvard University  
 MICHAEL L. MUSSA, University of Rochester  
 FRANK LEWIS, Queen's University
- 2:30 P.M. ECONOMIC ANALYSIS OF HOSPITALS\* (Joint Session with the Health Economics Research Organization)  
*Chairman:* DONALD E. YETT, University of Southern California  
*Papers:* KAREN DAVIS, The Brookings Institution  
 An Empirical Investigation of Alternative Models of the Hospital Industry  
 JOHN RAFFERTY, U.S. Department of Health, Education and Welfare  
 Net Prices and Non-Price Rationing: The Effects of Insurance on Hospital Output Mix  
 JUDITH LAVE, LESTER LAVE, and LESTER SILVERMAN, Carnegie-Mellon University  
 A Proposal for Incentive Reimbursement of Hospitals  
*Discussants:* PHEOBUS DHRAMES, University of California at Los Angeles and University of Southern California  
 HERBERT E. KLARMAN, New York University  
 MARK PAULY, Northwestern University

\* Not to appear in the *Papers and Proceedings*.

2:30 P.M. IMPLICIT GRANTS WITHIN THE NATIONAL BUDGET\* (Joint Session with the Association for the Study of the Grants Economy)

*Chairman:* MARTIN PFAFF, Wayne State University and University of Augsburg

*Papers:* MURRAY WEIDENBAUM, Washington University

Implicit Grants in the U.S. Budget: Some Recent Trends

PAUL SENF, University of the Saarland, F. R. of Germany

The Subsidy Report of the Federal Republic of Germany: The Treatment of Implicit Grants

JERRY JASINOWSKI, U.S. Senate Joint Economic Committee

The Need for the Reform of the U.S. Subsidy System

*Discussants:* BURKHARDT STRUMPEL, University of Michigan

THOMAS MULLER, The Urban Institute

(To be announced)

2:30 P.M. INTERNATIONAL TRADE

*Chairman:* ROBERT BALDWIN, University of Wisconsin

*Papers:* HELEN JUNZ, Division of International Finance, Federal Reserve System and RUDOLPH

R. RHOMBERG, International Monetary Fund

Price Competitiveness in Export Trade Among Industrial Countries

JAGDISH BHAGWATI, Massachusetts Institute of Technology AND ANNE KRUEGER, University of Minnesota

Exchange Control, Liberalization, and Economic Growth

*Discussants:* EDWARD E. LEAMER, Harvard University

GUSTAV RANIS, Yale University

# ANNOUNCEMENTS

## *Nominating Committee of The American Economic Association*

In accordance with Section IV, paragraph 2, of the bylaws of the American Economic Association as amended last year, President-elect Kenneth Arrow has appointed a Nominating Committee for 1972 consisting of Wassily Leontief, Chairman, Carolyn Shaw Bell, Leonid Hurwicz, Mark Killingsworth, Roy Radner, and David C. Smith. Attention of members is called to the part of the bylaw reading, "In addition to appointees chosen by the President-elect, the Committee shall include any other member of the Association nominated by petition including signatures and addresses of not less than 2 percent of the members of the Association, delivered to the Secretary before December 1. No member of the Association may validly petition for more than one nominee for the Committee. The names of the Committee shall be announced to the membership immediately following its appointment and the membership invited to suggest nominees for the various offices to the Committee."

## *Notices to Members and Subscribers*

Members of the American Economic Association receive preregistration materials for the annual meetings together with the Association's publications. Normally subscribers do not receive preregistration materials because most subscriptions are held by institutions. In some cases, however, individual subscribers may wish to receive preregistration materials. These may be obtained by writing to the secretary's office, Nashville, Tennessee.

The Asia Foundation has provided the American Economic Association with a fund to be used to assist students and visiting scholars from Asia who are studying in the United States or Canada to attend the annual meeting of the American Economic Association. The 1972 meeting will be held in Toronto, Canada, December 27-30. The maximum amount of money available to an individual applicant is \$150. Inquiries should be addressed to the American Economic Association, 1313 21st Avenue, South, Nashville, Tennessee 37212.

## *Case Studies Wanted in Manpower Policy and Manpower Administration*

In conjunction with a new program for training administrative staff in the manpower field, the John F. Kennedy School of Government is seeking to commission a series of case studies in manpower policy and its administration. These studies should be patterned after business school teaching cases and should provide students of manpower administration with illustrative problems in the areas of management decision making, planning, intergovernmental relations, and the like.

Case studies will receive informal publication by the Kennedy School (which will not, however, retain copyrights in material it uses). Cases should be short (less than 30 typewritten pages in all but the most exceptional cases) and self-contained. Case preparation fees start at \$400. Proposals for and drafts of cases to be included in the series should be sent to Professor Peter B. Doeringer, Harvard University, 1737 Cambridge St., Cambridge, Mass. 02138

TIMS Twentieth International Meeting will be held at Tel Aviv, Israel, June 24-29, 1973. The theme of the meeting will be the role of Management Sciences and allied disciplines in the achievement of social and economic goals in developing countries. Send abstracts of contributed papers to Professor Matthew J. Sobel, Department of Administrative Sciences, Yale University, New Haven, Connecticut 06520. The deadline for papers is Dec. 31, 1972.

The International Research and Exchanges Board, sponsored by the American Council of Learned Societies and the Social Science Research Council, administers academic exchanges with Eastern Europe and the Soviet Union. We have become increasingly aware of the need to encourage qualified scholars outside of the traditional area disciplines to participate in the research and exchange programs we offer. Our aim is to enable scholars in the humanities, social sciences and natural sciences to extend their principal disciplines to include study in that area and to take advantage of comparative materials in their research. The selection committees for the various programs are giving close attention to applications from non-area as well as area specialists, and a new Preparatory Fellowship program has been established to enable graduate students in underrepresented disciplines to acquire an area expertise prior to participation in the exchanges.

Further information on all of these programs, including eligibility requirements, duration of the grants, financial provisions and other details may be obtained by writing directly to the International Research and Exchanges Board, 110 East 59th Street, New York, New York 10022. In writing for information, please state your academic status, academic affiliation, field of specialization, and citizenship.

Omicron Delta Epsilon, the International Honor Society in Economics, invites the submission of entries for the fifth year of the Irving Fisher and Frank W. Taussig Award competitions. The Fisher Award consists of \$1,000 and publication of the manuscript as a monograph by the Princeton University Press, or a journal article to appear in the *American Economic Review*. The Taussig Award consists of \$100 and publication in *The American Economist*.

Entries should be submitted to the Departmental Selection Committees by January 1, 1973. They will be judged by the International Editorial Board and finalists by the Final Selection Board consisting of Professors Maurice Allais, William Baumol, Leonid Hurwicz, Wassily Leontief, and Egon Neuberger (editor). For more information, write to Egon Neuberger, Editor, Economic Research Bureau, State University of New York, Stony Brook, New York 11790.

#### *Scholarships for Study and Research in Germany*

The German Academic Exchange Service (DAAD) offers full scholarships to American students in all fields for graduate, doctoral, or postdoctoral studies at German institutions of higher learning during the academic year of 1973-74. Candidates must be between 18 and 32 years of age and have a good command of the German language. They must hold the Bachelor's degree at the time of the award and may hold the Ph.D., provided it was obtained no earlier than June 1970. A monthly allowance of DM 600 or DM 800, according to the candidate's academic level, will be granted for a period of 10 months, beginning October 1, 1973. The award includes round-trip transportation, waiver of tuition, incidentals allowance, baggage and book allowance, health insurance, dependent's allowance for married grantees, and, in some cases, a language course at a Goethe Institute. Application forms are obtainable from the Fulbright Program Advisor on the university campuses or the Institute of International Education (IIE), 809 United Nations Plaza, New York, New York 10017. Application deadline for 1973-74 grants is November 1, 1972 at the IIE.

#### *Professional Placement Service at the ASSA Annual Meeting, December 28-30, 1972*

The Department of Manpower and Immigration in cooperation with the Allied Social Science Associations is planning a professional placement service during the 1972 annual meetings in Toronto. The service will be located in the Four Seasons Sheraton Hotel, Grand Ballroom floor, and will operate from 9:00 A.M. to 6:00 P.M., December 28 and 29, and from 9:00 A.M. to 12:00 noon on December 30. For your convenience, the department is offering a pre-convention registration service. Listing your application or your employer orders prior to the convention will expedite service at the placement center. Applications and orders from individuals who do not plan to attend the convention will be accepted and made available for review.

Immediately upon arrival at the convention, registrants should report to the Convention Placement Center so that pre-convention registration can be activated. Professional placement order and application forms may be obtained by sending a request no later than November 30 to Department of Manpower and Immigration, Convention Placement Service, Box 23,

Toronto-Dominion Centre, Toronto 111, Ontario, Canada.

#### *Deaths*

Frank H. Knight, professor emeritus, University of Chicago, Apr. 15, 1972.

Warren L. Smith, professor of economics, University of Michigan, Apr. 23, 1972.

#### *Retirements*

Merwyn G. Bridenstine, department of economics, University of Arkansas, June 30, 1972.

George Katona, professor emeritus of economics and psychology, University of Michigan, June 1972.

#### *Visiting Foreign Scholars*

Frans Altling von Geusau, John F. Kennedy Institute, Tilburg, The Netherlands: visiting professor, Drew University, Sept. 1972.

N. Devletoglu, London School of Economics: visiting lecturer in economics, Virginia Polytechnic Institute and State University, spring 1973.

Celso Furtado, University of Paris: visiting professor of economics, NSF senior foreign scientist fellowship, American University, fall 1972.

Joseph K. Maitha, University of Nairobi: visiting research fellow, National Bureau of Economic Research, 1972.

James Meade, University of Cambridge: visiting professor of economics, Virginia Polytechnic Institute and State University, spring 1973.

Ryoichi Mikitani, Kobe University, Japan: visiting scholar, department of economics, Yale University, Sept. 1972-June 1973.

Owodunni Terriba, University of Ibadan: visiting associate professor and research associate, department of economics, University of Michigan, 1972-73.

#### *Promotions*

Ernst Baltensperger: associate professor of economics, Ohio State University.

Christopher Barnekov: assistant professor of economics, Ohio State University.

R. N. Batra: associate professor, University of Western Ontario, July 1, 1971.

Ira Brous: associate professor of economics, Ithaca College.

John W. Budina, Jr.: professor, department of economics and finance, Florida Technological University.

Michael P. Claudon: assistant professor, department of economics, Middlebury College, Feb. 1972.

Donald P. Cole: associate professor, department of economics, Drew University, Sept. 1, 1972.

Charles H. Cuykendall: associate professor and extension economist farm management, agricultural and applied economics and extension service, University of Minnesota.

Ralph H. Gelder: assistant vice president, bank supervision and relations, Federal Reserve Bank of New York.

Charles J. Goetz: professor, Virginia Polytechnic Institute and State University, Sept. 1972.

Margaret L. Greene: chief, financial statistics division, Federal Reserve Bank of New York.

Patric H. Hendershott: professor of economics and finance, Purdue University, Aug. 1972.

Bentti O. Hoiska: assistant professor of economics, Union College, Sept. 1972.

Ralph Kaminsky: professor of economics, Graduate School of Public Administration, New York University, Sept. 1, 1972.

Thomas R. Kershner: assistant professor of economics, Union College, Sept. 1972.

E. Dwayne Key: associate professor, department of economics, Stephen F. Austin State University, Sept. 1, 1972.

Bruce W. Kimzey: associate professor of economics, New Mexico State University, July 1, 1972.

Tetsunori Koizumi: associate professor of economics, Ohio State University.

Leon Korobow: manager, banking studies department, Federal Reserve Bank of New York.

Gene Laber: associate professor of economics, University of Vermont, Sept. 1, 1972.

Leonard Lapidus: vice president, public information, Federal Reserve Bank of New York.

J. C. Leith: associate professor, department of economics, University of Western Ontario, July 1, 1971.

Wilbur G. Lewellen: professor of industrial management, Purdue University, Aug. 1972.

Carl E. Liedholm: professor of economics, Michigan State University, July 1, 1972.

Charles M. Lucas: chief, market statistics division, Federal Reserve Bank of New York.

Charles E. Metcalf: associate professor of economics, University of Wisconsin-Madison.

Oscar Miller: professor, department of economics, University of Illinois at Chicago Circle.

Frank W. Musgrave: associate professor of economics, Ithaca College.

Willis L. Peterson: professor of agricultural and applied economics, University of Minnesota.

Edward Prescott: associate professor of economics, Graduate School of Industrial Administration, Carnegie-Mellon University, Sept. 1972.

James B. Ramsey: professor of economics, Michigan State University, July 1, 1972.

E. Saraydar: associate professor, department of economics, University of Western Ontario, July 1, 1972.

John T. Scott, Jr.: professor of agricultural economics, University of Illinois at Urbana-Champaign, spring 1972.

Allen Sinai: associate professor, department of economics, University of Illinois at Chicago Circle.

M. D. Stewart, Jr.: associate professor, department of economics, Stephen F. Austin State University, Sept. 1, 1972.

Michael Szenberg: associate professor of economics, Long Island University.

Emanuel Tobier: professor of economics, Graduate School of Public Administration, New York University, Sept. 1, 1972.

Joseph A. Wahed: assistant vice president and econo-

mist, economics department, Wells Fargo Bank, San Francisco, spring 1972.

Joseph P. Waters: assistant professor, department of economics, Middlebury College, Feb. 1972.

Warren E. Weber: associate professor, Virginia Polytechnic Institute and State University, Sept. 1972.

Delane E. Welsch: professor of agricultural and applied economics, University of Minnesota.

H. David Willey: assistant vice president, loans and credits, Federal Reserve Bank of New York.

### *Administrative Appointments*

Robert F. Dernberger: associate chairman, department of economics, University of Michigan.

Donald R. Gilmore: assistant vice president, National Bureau of Economic Research, Apr. 1, 1972.

Clyde A. Hauman, associate professor, College of William and Mary: director, Marshall-Wythe Institute for Research in the Social Sciences, July 1972.

Robert E. Hicks: chairman, department of economics and finance, Florida Technological University.

Lamar B. Jones: chairman, department of economics, Louisiana State University.

Phillip S. Kaplan: vice president for academic affairs and provost, University of New Haven, Apr. 1, 1972.

Warren C. Lackstrom: assistant vice president, National Bureau of Economic Research, Computer Research Center for Economics and Management Science.

Robert T. Michael: assistant vice president, National Bureau of Economic Research, July 1, 1972.

Joseph A. Parker: dean, Graduate School, University of New Haven, Apr. 1, 1972.

Robert C. Rose: chairman, department of economics, Saint Mary's College, Aug. 1, 1972.

Franklin B. Sherwood: chairman, department of economics, University of New Haven, Apr. 1, 1972.

Albert H. Small, U.S. Department of Commerce: assistant director, program operations, Price Commission, Apr. 1972.

Warren Smith: dean, School of Business Administration, University of New Haven, Apr. 1, 1972.

Ward Theilman: chairman, department of business administration, University of New Haven, Apr. 1, 1972.

Sheila Tschinkel: chief, securities analysis division, Federal Reserve Bank of New York.

James A. Zwerneman: professor, acting head, department of economics, New Mexico State University, Aug. 1972.

### *Appointments*

Biyan Aghevli, Brown University: economist, International Monetary Fund, Sept. 1972.

Swarnjit S. Arora, State University of New York at Buffalo: research fellow, National Bureau of Economic Research, Computer Research Center for Economics and Management Science, summer 1972.

Robert E. Baldwin: research associate, National Bureau of Economic Research.

Jere Behrman: research associate, National Bureau of Economic Research.

Richard Bernstein, Brown University: assistant professor of economics, Temple University, Sept. 1972.

Jagdish Bhagwati: research associate, National Bureau of Economic Research.

Helge K. Bjaaland: research associate, National Bureau of Economic Research, Computer Research Center for Economics and Management Science.

Raford Boddy, State University of New York at Buffalo: associate professor of economics, American University.

Carolyn Bomberger, Brown University: instructor in economics, Wayne State University, Sept. 1972.

William Bomberger, Brown University: assistant professor of economics, Wayne State University, Sept. 1972.

Eliyahu Borukhov, Johns Hopkins University: assistant professor, department of economics, Ohio State University.

Jeffrey I. Chapman, University of California, Berkeley: assistant research economist, Institute of Government and Public Affairs, University of California, Los Angeles, Sept. 1, 1971.

Robert J. Cline, University of Michigan: assistant professor of economics, Georgia State University, Sept. 1, 1972.

Scott F. Clow, University of Illinois: assistant professor of economics, California State College, Bakersfield, fall 1972.

William Courtney, Brown University: foreign service officer, State Department, Washington, D.C.

Charles C. Cox, University of Chicago: assistant professor, Ohio State University.

Jean David: assistant professor of economics, Florida Technological University.

Carlos Diaz-Alejandro: research associate, National Bureau of Economic Research.

Stanley Diller: research associate, National Bureau of Economic Research.

Larry Dixon: instructor of economics, Drew University, Sept. 1, 1972.

Walter Dolde, Yale University: research associate, Graduate School of Industrial Administration, Carnegie-Mellon University, Sept. 1972.

Vassilis Droucopoulos: lecturer, department of economics, Laurentian University, fall 1971.

S. Michael Edgar, University of Oklahoma: assistant professor of business administration, University of Texas at Arlington, July 15, 1971.

Mark Eisner: technical director, National Bureau of Economic Research, Computer Research Center for Economics and Management Science.

Frank Falero, Virginia Polytechnic Institute and State University: associate professor of economics, California State College, Bakersfield, fall 1972.

Donald E. Farrar: research associate, National Bureau of Economic Research.

Albert Fishlow: research associate, National Bureau of Economic Research.

Richard L. Floyd: assistant professor, department of economics and finance, University of Texas at El Paso, Sept. 1972.

William H. Foeller: assistant professor of economics, University of Akron.

Charles Frank: research associate, National Bureau of Economic Research.

Bernard Friedman, Massachusetts Institute of Technology: assistant professor of economics, Brown University, Sept. 1972.

Paul B. Ginsburg, Harvard University: assistant professor, department of economics, Michigan State University, Sept. 1972.

Lawrence Goldberg, Brown University: assistant professor of economics, Clark University, Sept. 1972.

Robert Goldberg: research associate, National Bureau of Economic Research.

Michael Gort: research associate, National Bureau of Economic Research.

Henry G. Grabowski: research associate, National Bureau of Economic Research.

Bent Hansen: research associate, National Bureau of Economic Research.

Tatsuo Hatta, Johns Hopkins University: assistant professor, Ohio State University.

James J. Heckman: research fellow, National Bureau of Economic Research, Sept. 1972.

David C. Hoaglin: research associate, National Bureau of Economic Research's Computer Research Center for Economics and Management Science.

William Holahan, Brown University: assistant professor of economics, University of Wisconsin-Milwaukee, Sept. 1972.

Paul W. Holland: senior research associate, National Bureau of Economic Research, Computer Research Center for Economics and Management Science, July 1, 1972.

Helen M. Hunter, Swarthmore College: associate professor, department of economics, Bryn Mawr College, Sept. 1972.

George R. Iden, University of North Carolina: associate professor of economics, American University.

Edward J. Kane, Boston College: professor of banking and monetary economics, Ohio State University.

John Kennen, Northwestern University: assistant professor of economics, Brown University, Sept. 1972.

James M. Kenney: instructor in economics, Union College, Sept. 1972.

Carolyn V. Kent: instructor in economics, Union College, Sept. 1972.

John Kirsch: support staff coordinator, National Bureau of Economic Research, Computer Research Center for Economics and Management Science.

Virginia C. Klema: research associate, National Bureau of Economic Research, Computer Research Center for Economics and Management Science.

Anne Krueger: research associate, National Bureau of Economic Research.

Edwin Kuh: director, National Bureau of Economic Research, Computer Research Center for Economics and Management Science.

Arleen Leibowitz: research associate, National Bureau of Economic Research, Feb. 1972.

Danny Leipziger, Brown University: international economist, Agency for International Development, Sept. 1972.

J. Clark Leith: research associate, National Bureau of Economic Research.

Robert A. Leone: research associate, National Bureau of Economic Research.

An-loh Lin: research associate, National Bureau of Economic Research.

Ben-chieh Liu, St. Louis Regional Industrial Development Corporation: senior economist, Midwest Research Institute, Kansas City, June 1, 1972.

Robert J. Mackay, University of North Carolina: assistant professor, Virginia Polytechnic Institute and State University.

Richard W. Nelson: economist, banking studies department, Federal Reserve Bank of New York.

William Orchard-Hays: senior research associate, National Bureau of Economic Research, Computer Research Center for Economics and Management Science, Jan. 1972.

Naomi Perlman: research associate, National Bureau of Economic Research.

E. Dwight Phaup: instructor in economics, Union College, Sept. 1972.

Richard A. Posner: research associate National Bureau of Economic Research.

Charlene Ramsey: instructor, department of economics, Florida Technological University, Mar. 1972.

Robert H. Rasche, University of Pennsylvania: associate professor, department of economics, Michigan State University, Sept. 1972.

Melvin W. Reder: research associate, National Bureau of Economic Research.

Martin Ringo, Brown University: assistant professor of economics, Miami University, Oxford, Sept. 1972.

Donald J. Rose: research associate, National Bureau of Economic Research, Computer Research Center for Economics and Management Science, Oct. 1972.

Daniel L. Rubinfeld: assistant professor, department of economics, University of Michigan.

F. M. Scherer, University of Michigan: senior research fellow, International Institute of Management, Berlin, Germany, Sept. 1972.

Richard Schuler, Brown University: assistant professor of economics, Cornell University, Sept. 1972.

William D. Schulze, University of California, Riverside: assistant professor of economics, University of New Mexico, Aug. 1972.

Steven A. Seelig: economist, banking studies department, Federal Reserve Bank of New York.

David Shapiro, Princeton University: assistant professor, department of economics, Ohio State University.

Heather Slemmer: assistant professor, department of economics, Florida Technological University, Sept. 1972.

Robert M. Spann, North Carolina State University: assistant professor, Virginia Polytechnic Institute and State University.

Wayne W. Snyder, University of Michigan: professor of economics, Sangamon State University, Apr. 1972.

T. N. Srinivasan: research associate, National Bureau of Economic Research.

Gilbert Suzawa, Brown University: assistant professor of economics, University of Rhode Island, Sept. 1972.

Lester D. Taylor: research associate, National Bureau of Economic Research.

Robert D. Tollen: assistant professor, department of economics and finance, University of Texas at El Paso, Sept. 1972.

Allen Vandermeulen, Brown University: assistant professor of economics, Dartmouth College, Sept. 1972.

Paul Wachtel: research associate, National Bureau of Economic Research.

Susan M. Wachter: lecturer, department of economics, Bryn Mawr College, Sept. 1972.

Thomas E. Weisskopf, Harvard University: associate professor, department of economics, University of Michigan.

Gavin Wright, Yale University: assistant professor, department of economics, University of Michigan.

### *Leaves for Special Appointments*

John Brown, Brown University: fellow in law in economics, University of Chicago Law School, 1972-73.

Jan W. Duggar, Louisiana State University: scientific director, Gulf South Research Institute.

Victor R. Fuchs, National Bureau of Economic Research: fellow, Center for Advanced Study in the Behavioral Sciences, Sept. 1972-Aug. 1973.

Samuel Gubins, Haverford College: director of research, Center for National Policy Review, 1971-72.

Richard D. Porter, Ohio State University: Board of Governors of Federal Reserve Bank, Washington, D.C.

Ronald Savitt, Boston University: faculty of business administration and commerce, University of Alberta, Jan. 1, 1973; Fulbright lectureship, Bogazici University, Istanbul, Turkey, 1972-73.

Seymour I. Somberg, Stephen F. Austin State University: FAO/UN consultant to government of Surinam.

Jerome Stein, Brown University: visiting professor of economics, Guggenheim fellowship, Hebrew University, Israel, 1972-73.

John C. Weicher, associate professor of economics, Ohio State University: staff economist, Office of Economic Opportunity, Washington, D.C.

Burton A. Weisbrod, University of Wisconsin-Madison: visiting professor of economic policy, State University, Binghamton, fall semester 1972.

Frank C. Wykoff, Pomona College: Brookings Institution, June 5, 1972.

### *Resignations*

Martin David, University of Michigan: University of Wisconsin, June 1972.

Solomon Fabricant, National Bureau of Economic Research, June 30, 1972.

David E. Lindsey, Ohio State University: Macalester College, June 1972.

Norman C. Miller, Carnegie-Mellon University: University of Pittsburgh, Sept. 1972.

Richard T. Stillson, Ohio State University: International Monetary Fund, Sept. 1972.

John Stone, Federal Reserve Bank of New York: Ford Foundation.

Douglas W. Webbink, University of North Carolina at Chapel Hill: Federal Trade Commission, Bureau of Economics, Aug. 1972.



# THE AMERICAN ECONOMIC ASSOCIATION

Founded in 1885

• Published at George Banta Co., Inc., Menasha, Wisconsin.

• THE AMERICAN ECONOMIC REVIEW, including four quarterly numbers, the *Proceedings* of the annual meetings, and *Directory* and Supplements, is published by the American Economic Association and is sent to all members five times a year, in March, May, June, September, and December.

• Membership dues of the Association are \$20.00 a year, which includes a year's subscription to both the *American Economic Review* and the *Journal of Economic Literature*. Subscriptions by nonmembers are \$30.00 a year, and only subscriptions to both publications will be accepted. Single copies of the *Review* and *Journal* are \$4.00 each. Each order for copies of either publication must also include a \$.50 per order service charge. Orders should be sent to the Secretary's office, Nashville, Tennessee.

• Correspondence relating to the *Papers and Proceedings*, the *Directory*, advertising, permission to quote, business matters, subscriptions, membership and changes of address may be sent to the secretary, Rendigs Fels, 1313 21st Avenue, South, Nashville, Tennessee 37212. To be effective, notice of change of address must reach the secretary by the 1st of the month previous to the month of publication. The Association's publications are mailed by second class and are not forwardable by the Post Office.

• Second-class postage paid at Nashville, Tennessee and at additional mailing offices. Printed in U.S.A.

## Officers

### *President*

JOHN KENNETH GALBRAITH  
Harvard University

### *President-Elect*

KENNETH ARROW  
Harvard University

### *Vice-Presidents*

HENDRIK S. HOUTHAKKER  
Harvard University  
ARTHUR M. OKUN  
Brookings Institution

### *Secretary-Treasurer and Editor of Proceedings*

RENDIGS FELS  
Vanderbilt University

### *Managing Editor of The American Economic Review*

GEORGE H. BORTS  
Brown University

### *Managing Editor of The Journal of Economic Literature*

MARK PERLMAN  
University of Pittsburgh

## Executive Committee

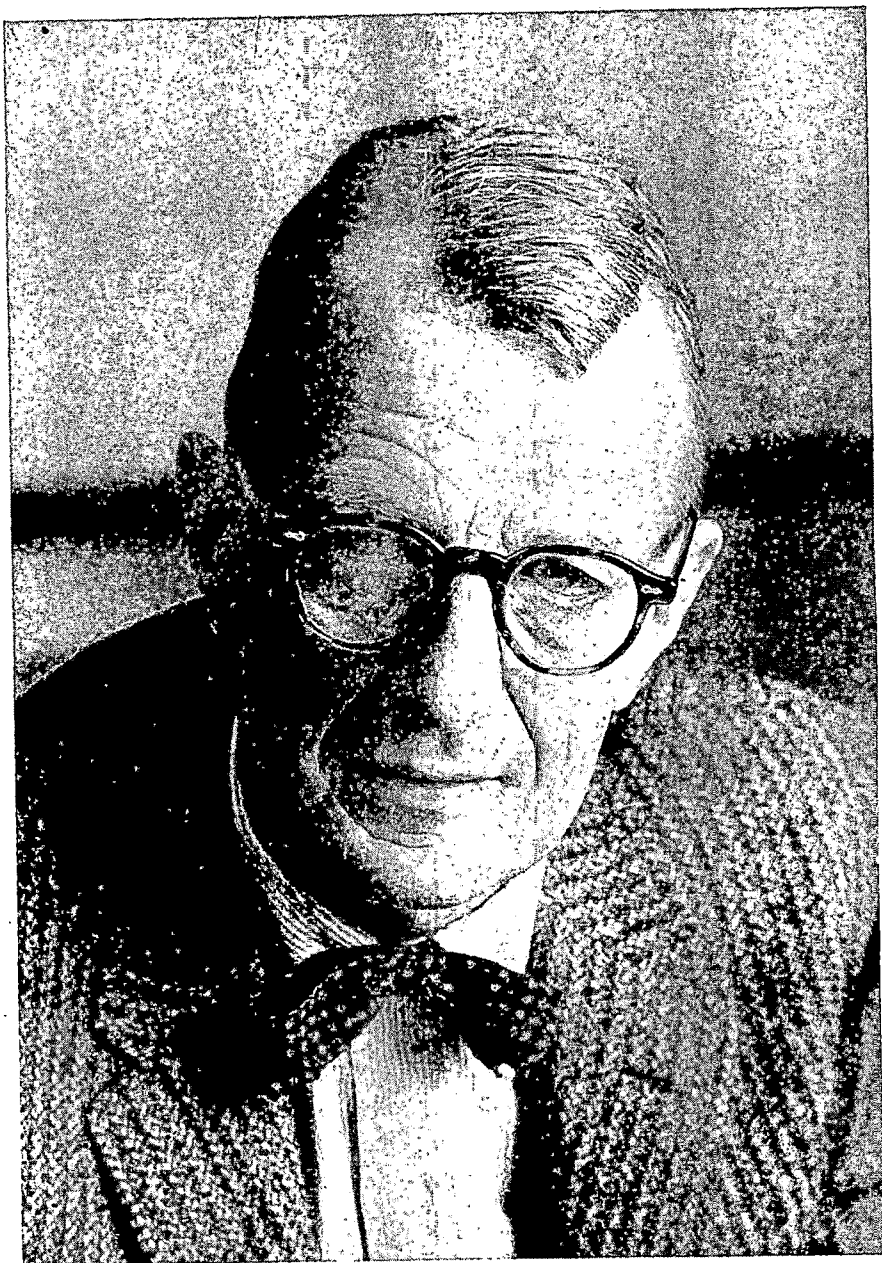
### *Elected Members of the Executive Committee*

ROBERT DORFMAN  
Harvard University  
ARNOLD C. HARBERGER  
University of Chicago  
ROBERT EISNER  
Northwestern University  
JOHN R. MEYER  
Yale University  
GUY HENDERSON ORCUTT  
Yale University  
JOSEPH A. PECHMAN  
Brookings Institution

### *Ex Officio Members*

WASSILY LEONTIEF  
Harvard University  
JAMES TOBIN  
Yale University

ARTHUR SMITHIES  
Editor, *Journal of Economic Abstracts*  
1963-68



Arthur Smithers

# Production, Information Costs, and Economic Organization

By ARMEN A. ALCHIAN AND HAROLD DEMSETZ\*

The mark of a capitalistic society is that resources are owned and allocated by such nongovernmental organizations as firms, households, and markets. Resource owners increase productivity through cooperative specialization and this leads to the demand for economic organizations which facilitate cooperation. When a lumber mill employs a cabinetmaker, cooperation between specialists is achieved within a firm, and when a cabinetmaker purchases wood from a lumberman, the cooperation takes place across markets (or between firms). Two important problems face a theory of economic organization—to explain the conditions that determine whether the gains from specialization and cooperative production can better be obtained within an organization like the firm, or across markets, and to explain the structure of the organization.

It is common to see the firm characterized by the power to settle issues by fiat, by authority, or by disciplinary action superior to that available in the conventional market. This is delusion. The firm does not own all its inputs. It has no power of fiat, no authority, no disciplinary action any different in the slightest degree from ordinary market contracting between any two people. I can "punish" you only by withholding future business or by seeking redress in the courts for any failure to honor our exchange agreement. That is exactly all that any employer can do. He

can fire or sue, just as I can fire my grocer by stopping purchases from him or sue him for delivering faulty products. What then is the content of the presumed power to manage and assign workers to various tasks? Exactly the same as one little consumer's power to manage and assign his grocer to various tasks. The single consumer can assign his grocer to the task of obtaining whatever the customer can induce the grocer to provide at a price acceptable to both parties. That is precisely all that an employer can do to an employee. To speak of managing, directing, or assigning workers to various tasks is a deceptive way of noting that the employer continually is involved in renegotiation of contracts on terms that must be acceptable to both parties. Telling an employee to type this letter rather than to file that document is like my telling a grocer to sell me this brand of tuna rather than that brand of bread. I have no contract to continue to purchase from the grocer and neither the employer nor the employee is bound by any contractual obligations to continue their relationship. Long-term contracts between employer and employee are not the essence of the organization we call a firm. My grocer can count on my returning day after day and purchasing his services and goods even with the prices not always marked on the goods—because I know what they are—and he adapts his activity to conform to my directions to him as to what I want each day . . . he is not my employee.

Wherein then is the relationship between a grocer and his employee different from that between a grocer and his cus-

\* Professors of economics at the University of California, Los Angeles. Acknowledgment is made for financial aid from the E. Lilly Endowment, Inc. grant to UCLA for research in the behavioral effects of property rights.

tomers? It is in a *team* use of inputs and a centralized position of some party in the contractual arrangements of *all* other inputs. It is the *centralized contractual agent in a team productive process*—not some superior authoritarian directive or disciplinary power. Exactly what is a team process and why does it induce the contractual form, called the firm? These problems motivate the inquiry of this paper.

### I. The Metering Problem

The economic organization through which input owners cooperate will make better use of their comparative advantages to the extent that it facilitates the payment of rewards in accord with productivity. If rewards were random, and without regard to productive effort, no incentive to productive effort would be provided by the organization; and if rewards were negatively correlated with productivity the organization would be subject to sabotage. Two key demands are placed on an economic organization—metering input productivity and metering rewards.<sup>1</sup>

Metering problems sometimes can be resolved well through the exchange of products across competitive markets, because in many situations markets yield a high correlation between rewards and productivity. If a farmer increases his output of wheat by 10 percent at the prevailing market price, his receipts also increase by 10 percent. This method of organizing economic activity meters the *output directly*, reveals the marginal product and apportions the *rewards* to resource owners in accord with that direct measurement of their outputs. The success of this decentralized, market exchange in promoting productive specialization requires that changes in market rewards fall

on those responsible for changes in *output*.<sup>2</sup>

The classic relationship in economics that runs from marginal productivity to the distribution of income implicitly *assumes* the existence of an organization, be it the market or the firm, that allocates rewards to resources in accord with their productivity. The problem of economic organization, the economical means of metering productivity and rewards, is not confronted directly in the classical analysis of production and distribution. Instead, that analysis tends to assume sufficiently economic—or zero cost—means, as if productivity automatically created its reward. We conjecture the direction of causation is the reverse—the specific sys-

<sup>2</sup> A producer's wealth would be reduced by the present capitalized value of the future income lost by loss of reputation. Reputation, i.e., credibility, is an asset, which is another way of saying that reliable information about expected performance is both a costly and a valuable good. For acts of God that interfere with contract performance, both parties have incentives to reach a settlement akin to that which would have been reached if such events had been covered by specific contingency clauses. The reason, again, is that a reputation for "honest" dealings—i.e., for actions similar to those that would probably have been reached had the contract provided this contingency—is wealth.

Almost every contract is open-ended in that many contingencies are uncovered. For example, if a fire delays production of a promised product by *A* to *B*, and if *B* contends that *A* has not fulfilled the contract, how is the dispute settled and what recompense, if any, does *A* grant to *B*? A person uninitiated in such questions may be surprised by the extent to which contracts permit either party to escape performance or to nullify the contract. In fact, it is hard to imagine any contract, which, when taken solely in terms of its stipulations, could not be evaded by one of the parties. Yet that is the ruling, viable type of contract. Why? Undoubtedly the best discussion that we have seen on this question is by Stewart Macaulay.

There are means not only of detecting or preventing cheating, but also for deciding how to allocate the losses or gains of unpredictable events or quality of items exchanged. Sales contracts contain warranties, guarantees, collateral, return privileges and penalty clauses for specific nonperformance. These are means of assignment of *risks* of losses of cheating. A lower price without warranty—an "as is" purchase—places more of the risk on the buyer while the seller buys insurance against losses of his "cheating." On the other hand, a warranty or return privilege or service contract places more risk on the seller with insurance being bought by the buyer.

<sup>1</sup> Meter means to measure and also to apportion. One can meter (measure) output and one can also meter (control) the output. We use the word to denote both; the context should indicate which.

tem of rewarding which is relied upon stimulates a particular productivity response. If the economic organization meters poorly, with rewards and productivity only loosely correlated, then productivity will be smaller; but if the economic organization meters well productivity will be greater. What makes metering difficult and hence induces means of economizing on metering costs?

## II. Team Production

Two men jointly lift heavy cargo into trucks. Solely by observing the total weight loaded per day, it is impossible to determine each person's marginal productivity. With team production it is difficult, solely by observing total output, to either define or determine *each* individual's contribution to this output of the cooperating inputs. The output is yielded by a team, by definition, and it is not a *sum* of separable outputs of each of its members. Team production of  $Z$  involves at least two inputs,  $X_i$  and  $X_j$ , with  $\partial^2 Z / \partial X_i \partial X_j \neq 0$ .<sup>3</sup> The production function is *not* separable into two functions each involving only inputs  $X_i$  or only inputs  $X_j$ . Consequently there is no *sum* of  $Z$  of two separable functions to treat as the  $Z$  of the team production function. (An example of a *separable* case is  $Z = aX_i^2 + bX_j^2$  which is separable into  $Z_i = aX_i^2$  and  $Z_j = bX_j^2$ , and  $Z = Z_i + Z_j$ . This is not team production.) There exist production techniques in which the  $Z$  obtained is greater than if  $X_i$  and  $X_j$  had produced separable  $Z$ . Team production will be used if it yields an output enough larger than the sum of separable production of  $Z$  to cover the costs of organizing and disciplining team members—the topics of this paper.<sup>4</sup>

<sup>3</sup> The function is separable into additive functions if the cross partial derivative is zero, i.e., if  $\partial^2 Z / \partial X_i \partial X_j = 0$ .

<sup>4</sup> With sufficient generality of notation and conception this team production function could be formulated as a case of the generalized production function interpretation given by our colleague, E. A. Thompson.

Usual explanations of the gains from cooperative behavior rely on exchange and production in accord with the comparative advantage specialization principle with separable additive production. However, as suggested above there is a source of gain from cooperative activity involving working as a *team*, wherein individual cooperating inputs do not yield identifiable, separate products which can be *summed* to measure the total output. For this cooperative productive activity, here called "team" production, measuring *marginal* productivity and making payments in accord therewith is more expensive by an order of magnitude than for separable production functions.

Team production, to repeat, is production in which 1) several types of resources are used and 2) the product is not a sum of separable outputs of each cooperating resource. An additional factor creates a team organization problem—3) not all resources used in team production belong to one person.

We do not inquire into why all the jointly used resources are not owned by one person, but instead into the types of organization, contracts, and informational and payment procedures used among owners of teamed inputs. With respect to the one-owner case, perhaps it is sufficient merely to note that (a) slavery is prohibited, (b) one might assume risk aversion as a reason for one person's not borrowing enough to purchase all the assets or sources of services rather than renting them, and (c) the purchase-resale spread may be so large that costs of short-term ownership exceed rental costs. Our problem is viewed basically as one of organization among different people, not of the physical goods or services, however much there must be selection and choice of combination of the latter.

How can the members of a team be rewarded and induced to work efficiently?

In team production, marginal products of cooperative team members are not so directly and separably (i.e., cheaply) observable. What a team offers to the market can be taken as the marginal product of the team but not of the team members. The costs of metering or ascertaining the marginal products of the team's members is what calls forth new organizations and procedures. Clues to each input's productivity can be secured by observing *behavior* of individual inputs. When lifting cargo into the truck, how rapidly does a man move to the next piece to be loaded, how many cigarette breaks does he take, does the item being lifted tilt downward toward his side?

If detecting such behavior were costless, neither party would have an incentive to shirk, because neither could impose the cost of his shirking on the other (if their cooperation was agreed to voluntarily). But since costs must be incurred to monitor each other, each input owner will have more incentive to shirk when he works as part of a team, than if his performance could be monitored easily or if he did not work as a team. If there is a net increase in productivity available by team production, net of the metering cost associated with disciplining the team, then team production will be relied upon rather than a multitude of bilateral exchange of separable individual outputs.

Both leisure and higher income enter a person's utility function.<sup>5</sup> Hence, each person should adjust his work and realized reward so as to equate the marginal rate of substitution between leisure and production of real output to his marginal rate of substitution in consumption. That is, he would adjust his rate of work to bring his demand prices of leisure and output to equality with their true costs. However,

with detection, policing, monitoring, measuring or metering costs, each person will be induced to take more leisure, because the effect of relaxing on *his realized* (reward) rate of substitution between output and leisure will be less than the effect on the *true* rate of substitution. His realized cost of leisure will fall more than the true cost of leisure, so he "buys" more leisure (i.e., more nonpecuniary reward).

If his relaxation cannot be detected perfectly at zero cost, part of its effects will be borne by others in the team, thus making *his* realized cost of relaxation less than the true total cost to the team. The difficulty of detecting such actions permits the private costs of his actions to be less than their full costs. Since each person responds to his private realizable rate of substitution (in production) rather than the true total (i.e., social) rate, and so long as there are costs for other people to detect his shift toward relaxation, it will not pay (them) to force him to readjust completely by making him realize the true cost. Only enough efforts will be made to equate the marginal gains of detection activity with the marginal costs of detection; and that implies a lower rate of productive effort and more shirking than in a costless monitoring, or measuring, world.

In a university, the faculty use office telephones, paper, and mail for personal uses beyond strict university productivity. The university administrators could stop such practices by identifying *the* responsible person in each case, but they can do so only at higher costs than administrators are willing to incur. The extra costs of identifying each party (rather than merely identifying the presence of such activity) would exceed the savings from diminished faculty "turpitudinal peccadilloes." So the faculty is allowed some degree of "privileges, perquisites, or fringe benefits." And the total of the pecuniary wages paid

<sup>5</sup> More precisely: "if anything other than pecuniary income enters his utility function." Leisure stands for all nonpecuniary income for simplicity of exposition.

is lower because of this irreducible (at acceptable costs) degree of amenity-seizing activity. Pay is lower in pecuniary terms and higher in leisure, conveniences, and ease of work. But still every person would prefer to see detection made more effective (if it were somehow possible to monitor costlessly) so that he, as part of the now more effectively producing team, could thereby realize a higher pecuniary pay and less leisure. If everyone could, at zero cost, have his reward-realized rate brought to the true production possibility real rate, all could achieve a more preferred position. But detection of the responsible parties is costly; that cost acts like a tax on work rewards.<sup>6</sup> Viable shirking is the result.

What forms of organizing team production will lower the cost of detecting "performance" (i.e., marginal productivity) and bring personally realized rates of substitution closer to true rates of substitution? Market competition, in principle, could monitor some team production. (It already *organizes* teams.) Input owners who are not team members can offer, in return for a smaller share of the team's rewards, to replace excessively (i.e., overpaid) shirking members. Market competition among potential team members would determine team membership and individual rewards. There would be no team leader, manager, organizer, owner, or employer. For such decentralized organizational control to work, outsiders, possibly after observing each team's total

output, can speculate about their capabilities as team members and, by a market competitive process, revised teams with greater productive ability will be formed and sustained. Incumbent members will be constrained by threats of replacement by outsiders offering services for lower reward shares or offering greater rewards to the other members of the team. Any team member who shirked in the expectation that the reduced output effect would not be attributed to him will be displaced if his activity is detected. Teams of productive inputs, like business units, would evolve in apparent spontaneity in the market—without any central organizing agent, team manager, or boss.

But completely effective control cannot be expected from individualized market competition for two reasons. First, for this competition to be completely effective, new challengers for team membership must know where, and to what extent, shirking is a serious problem, i.e., know they can increase net output as compared with the inputs they replace. To the extent that this is true it is probably possible for existing fellow team members to recognize the shirking. But, by definition, the detection of shirking by observing team output is costly for team production. Secondly, assume the presence of detection costs, and assume that in order to secure a place on the team a new input owner must accept a smaller share of rewards (or a promise to produce more). Then his incentive to shirk would still be at least as great as the incentives of the inputs replaced, because he still bears less than the entire reduction in team output for which he is responsible.

### III. The Classical Firm

One method of reducing shirking is for someone to specialize as a monitor to check the input performance of team members.<sup>7</sup>

<sup>6</sup> Do not assume that the sole result of the cost of detecting shirking is one form of payment (more leisure and less take home money). With several members of the team, each has an incentive to cheat against each other by engaging in more than the average amount of such leisure if the employer can not tell at zero cost which employee is taking more than average. As a result the total productivity of the team is lowered. Shirking detection costs thus change the form of payment and also result in lower total rewards. Because the cross partial derivatives are positive, shirking reduces other people's marginal products.

<sup>7</sup> What is meant by performance? Input energy, initiative, work attitude, perspiration, rate of exhaustion?

(Continued)

But who will monitor the monitor? One constraint on the monitor is the aforesaid market competition offered by other monitors, but for reasons already given, that is not perfectly effective. Another constraint can be imposed on the monitor: give him title to the net earnings of the team, net of payments to other inputs. If owners of cooperating inputs agree with the monitor that he is to receive any residual product above prescribed amounts (hopefully, the marginal value products of the other inputs), the monitor will have an added incentive not to shirk as a monitor. Specialization in monitoring plus reliance on a residual claimant status will reduce shirking; but additional links are needed to forge the firm of classical economic theory. How will the residual claimant monitor the other inputs?

We use the term monitor to connote several activities in addition to its disciplinary connotation. It connotes measuring output performance, apportioning rewards, observing the input behavior of inputs as means of detecting or estimating their marginal productivity and giving assignments or instructions in what to do and how to do it. (It also includes, as we shall show later, authority to terminate or revise contracts.) Perhaps the contrast between a football coach and team captain is helpful. The coach selects strategies and tactics and sends in instructions about what plays to utilize. The captain is essentially an observer and reporter of

the performance at close hand of the members. The latter is an inspector-steward and the former a supervisor manager. For the present all these activities are included in the rubric "monitoring." All these tasks are, in principle, negotiable across markets, but we are presuming that such market measurement of marginal productivities and job reassignments are not so cheaply performed for team production. And in particular our analysis suggests that it is not so much the costs of spontaneously negotiating contracts in the markets among groups for team production as it is the detection of the performance of individual members of the team that calls for the organization noted here.

The specialist *who receives the residual rewards* will be the monitor of the members of the team (i.e., will manage the use of cooperative inputs). The monitor earns his residual through the reduction in shirking that he brings about, not only by the prices that he agrees to pay the owners of the inputs, but also by observing and directing the actions or uses of these inputs. *Managing or examining the ways to which inputs are used in team production is a method of metering the marginal productivity of individual inputs to the team's output.*

To discipline team members and reduce shirking, the residual claimant must have power to revise the contract terms and incentives of *individual* members without having to terminate or alter every other input's contract. Hence, team members who seek to increase their productivity will assign to the monitor not only the residual claimant right but also the right to alter individual membership and performance on the team. Each team member, of course, can terminate his own membership (i.e., quit the team), but only the monitor may unilaterally terminate the membership of any of the

---

Or output? It is the latter that is sought—the *effect* or output. But performance is nicely ambiguous because it suggests both input and output. It is *nicely* ambiguous because as we shall see, sometimes by inspecting a team member's input activity we can better judge his output effect, perhaps not with complete accuracy but better than by watching the output of the *team*. It is not always the case that watching input activity is the only or best means of detecting, measuring or monitoring output effects of each team member, but in some cases it is a useful way. For the moment the word performance glosses over these aspects and facilitates concentration on other issues.

other members without necessarily terminating the team itself or his association with the team; and he alone can expand or reduce membership, alter the mix of membership, or sell the right to be the residual claimant-monitor of the team. It is this entire bundle of rights: 1) to be a residual claimant; 2) to observe input behavior; 3) to be the central party common to all contracts with inputs; 4) to alter the membership of the team; and 5) to sell these rights, that defines the *ownership* (or the employer) of the *classical* (capitalist, free-enterprise) firm. The coalescing of these rights has arisen, our analysis asserts, because it resolves the shirking-information problem of team production better than does the noncentralized contractual arrangement.

The relationship of each team member to the *owner* of the firm (i.e., the party common to all input contracts and the residual claimant) is simply a "quid pro quo" contract. Each makes a purchase and sale. The employee "orders" the owner of the team to pay him money in the same sense that the employer directs the team member to perform certain acts. The employee can terminate the contract as readily as can the employer, and long-term contracts, therefore, are not an essential attribute of the firm. Nor are "authoritarian," "dictatorial," or "fiat" attributes relevant to the conception of the firm or its efficiency.

In summary, two necessary conditions exist for the emergence of the firm on the prior assumption that more than pecuniary wealth enter utility functions: 1) It is possible to increase productivity through team-oriented production, a production technique for which it is costly to directly measure the marginal outputs of the co-operating inputs. This makes it more difficult to restrict shirking through simple market exchange between cooperating inputs. 2) It is economical to estimate mar-

ginal productivity by observing or specifying input behavior. The simultaneous occurrence of both these preconditions leads to the contractual organization of inputs, known as the *classical capitalist firms* with (a) joint input production, (b) several input owners, (c) one party who is common to all the contracts of the joint inputs, (d) who has rights to renegotiate any input's contract independently of contracts with other input owners, (e) who holds the residual claim, and (f) who has the right to sell his central contractual residual status.<sup>8</sup>

### *Other Theories of the Firm*

At this juncture, as an aside, we briefly place this theory of the firm in the contexts of those offered by Ronald Coase and Frank Knight.<sup>9</sup> Our view of the firm is not necessarily inconsistent with Coase's; we attempt to go further and identify refutable implications. Coase's penetrating insight is to make more of the fact that markets do not operate costlessly, and he relies on the cost of using markets to *form* contracts as his basic explanation for the existence of firms. We do not disagree with the proposition that, *ceteris paribus*, the higher is the cost of transacting across markets the greater will be the comparative advantage of organizing resources within the firm; it is a difficult proposition to disagree with or to refute. We could with equal ease subscribe to a theory of the firm based on the cost of managing, for surely it is true that, *ceteris paribus*, the lower is the cost of managing the greater will be the comparative advantage of organizing resources within the firm. To move the theory forward, it is necessary to know what is meant by a firm and to

<sup>8</sup> Removal of (b) converts a capitalist proprietary firm to a socialist firm.

<sup>9</sup> Recognition must also be made to the seminal inquiries by Morris Silver and Richard Auster, and by H. B. Malmgren.

explain the circumstances under which the cost of "managing" resources is low relative to the cost of allocating resources through market transaction. The conception of and rationale for the classical firm that we propose takes a step down the path pointed out by Coase toward that goal. Consideration of team production, team organization, difficulty in metering outputs, and the problem of shirking are important to our explanation but, so far as we can ascertain, not in Coase's. Coase's analysis insofar as it had heretofore been developed would suggest open-ended contracts but does not appear to imply anything more—neither the residual claimant status nor the distinction between employee and subcontractor status (nor any of the implications indicated below). And it is not true that employees are generally employed on the basis of long-term contractual arrangements any more than on a series of short-term or indefinite length contracts.

The importance of our proposed additional elements is revealed, for example, by the explanation of why the person to whom the control monitor is responsible receives the residual, and also by our later discussion of the implications about the corporation, partnerships, and profit sharing. These alternative forms for organization of the firm are difficult to resolve on the basis of market transaction costs only. Our exposition also suggests a definition of the classical firm—something crucial that was heretofore absent.

In addition, sometimes a technological development will lower the cost of market transactions while, at the same time, it expands the role of the firm. When the "putting out" system was used for weaving, inputs were organized largely through market negotiations. With the development of efficient central sources of power, it became economical to perform weaving in proximity to the power source and to engage in team production. The bringing

in of weavers surely must have resulted in a reduction in the cost of negotiating (forming) contracts. Yet, what we observe is the beginning of the factory system in which inputs are organized within a firm. Why? The weavers did not simply move to a common source of power that they could tap like an electric line, purchasing power while they used their own equipment. Now team production in the joint use of equipment became more important. The measurement of marginal productivity, which now involved interactions between workers, especially through their joint use of machines, became more difficult though contract negotiating cost was reduced, while managing the *behavior* of inputs became easier because of the increased centralization of activity. The firm as an organization expanded even though the cost of transactions was reduced by the advent of centralized power. The same could be said for modern assembly lines. Hence the emergence of central power sources expanded the scope of productive activity in which the firm enjoyed a comparative advantage as an organizational form.

Some economists, following Knight, have identified the bearing of risks of wealth changes with the director or central employer without explaining why that is a viable arrangement. Presumably, the more risk-averse inputs become employees rather than owners of the classical firm. Risk averseness and uncertainty *with regard to the firm's fortunes* have little, if anything, to do with our explanation although it helps to explain why all resources in a team are not owned by one person. That is, the role of risk taken in the sense of absorbing the windfalls that buffet the firm because of unforeseen competition, technological change, or fluctuations in demand are not central to our theory, although it is true that imperfect knowledge and, therefore, risk, in *this* sense of risk, underlie the problem of

monitoring team behavior. We deduce the system of paying the manager with a residual claim (the equity) from the desire to have efficient means to reduce shirking so as to make team production economical and not from the smaller aversion to the risks of enterprise in a dynamic economy. We conjecture that "distribution-of-risk" is not a valid rationale for the *existence* and organization of the *classical* firm.

Although we have emphasized team production as creating a costly metering task and have treated team production as an essential (necessary?) condition for the firm, would not other obstacles to cheap metering also call forth the same kind of contractual arrangement here denoted as a firm? For example, suppose a farmer produces wheat in an easily ascertained quantity but with subtle and difficult to detect quality variations determined by how the farmer grew the wheat. A vertical integration could allow a purchaser to control the farmer's behavior in order to more economically estimate productivity. But this is not a case of joint or team production, unless "information" can be considered part of the product. (While a good case could be made for that broader conception of production, we shall ignore it here.) Instead of forming a firm, a buyer can contract to have his inspector on the site of production, just as home builders contract with architects to supervise building contracts; that arrangement is not a firm. Still, a firm might be organized in the production of many products wherein no team production or jointness of use of separately owned resources is involved.

This possibility rather clearly indicates a broader, or complementary, approach to that which we have chosen. 1) As we do in this paper, it can be argued that the firm is the particular policing device utilized when joint team production is present. If other sources of high policing costs arise, as in the wheat case just indicated, some other form of contractual ar-

rangement will be used. Thus to each source of informational cost there may be a different type of policing and contractual arrangement. 2) On the other hand, one can say that where policing is difficult across markets, various forms of contractual arrangements are devised, but there is no reason for that known as the firm to be uniquely related or even highly correlated with team production, as defined here. It might be used equally probably and viably for other sources of high policing cost. We have not intensively analyzed other sources, and we can only note that our current and readily revisable conjecture is that 1) is valid, and has motivated us in our current endeavor. In any event, the test of the theory advanced here is to see whether the conditions we have identified are necessary for firms to have long-run viability rather than merely births with high infant mortality. Conglomerate firms or collections of separate production agencies into one owning organization can be interpreted as an investment trust or investment diversification device—probably along the lines that motivated Knight's interpretation. A holding company can be called a firm, because of the common association of the word firm with any ownership unit that owns income sources. The term firm as commonly used is so turgid of meaning that we can not hope to explain every entity to which the name is attached in common or even technical literature. Instead, we seek to identify and explain a particular contractual arrangement induced by the cost of information factors analyzed in this paper.

#### IV. Types of Firms

##### A. Profit-Sharing Firms

Explicit in our explanation of the capitalist firm is the assumption that the cost of *managing* the team's inputs by a central monitor, who disciplines himself because he is a residual claimant, is low

relative to the cost of metering the marginal outputs of team members.

If we look within a firm to see who monitors—hires, fires, changes, promotes, and renegotiates—we should find him being a residual claimant or, at least, one whose pay or reward is more than any others correlated with fluctuations in the residual value of the firm. They more likely will have options or rights or bonuses than will inputs with other tasks.

An implicit “auxiliary” assumption of our explanation of the firm is that the cost of team production is increased if the residual claim is not held entirely by the central monitor. That is, we assume that if profit sharing had to be relied upon for *all* team members, losses from the resulting increase in central monitor shirking would exceed the output gains from the increased incentives of other team members not to shirk. If the optimal team size is only two owners of inputs, then an equal division of profits and losses between them will leave each with stronger incentives to reduce shirking than if the optimal team size is large, for in the latter case only a smaller percentage of the losses occasioned by the shirker will be borne by him. Incentives to shirk are positively related to the optimal size of the team under an equal profit-sharing scheme.<sup>10</sup>

The preceding does not imply that profit sharing is never viable. Profit sharing to encourage self-policing is more appropriate for small teams. And, indeed, where input owners are free to make whatever contractual arrangements suit them, as generally is true in capitalist economies, profit sharing seems largely limited to partner-

ships with a relatively small number of *active*<sup>11</sup> partners. Another advantage of such arrangements for smaller teams is that it permits more effective reciprocal monitoring among inputs. Monitoring need not be entirely specialized.

Profit sharing is more viable if small team size is associated with situations where the cost of specialized management of inputs is large relative to the increased productivity potential in team effort. We conjecture that the cost of managing team inputs increases if the productivity of a team member is difficult to correlate with his behavior. In “artistic” or “professional” work, watching a man’s activities is not a good clue to what he is actually thinking or doing with his mind. While it is relatively easy to manage or direct the loading of trucks by a team of dock workers where input activity is so highly related in an obvious way to output, it is more difficult to manage and direct a lawyer in the preparation and presentation of a case. Dock workers can be directed in detail without the monitor himself loading the truck, and assembly line workers can be monitored by varying the speed of the assembly line, but detailed direction in the preparation of a law case would require in much greater degree that the monitor prepare the case himself. As a result, artistic or professional inputs, such as lawyers, advertising specialists, and doctors, will be given relatively freer reign with regard to individual behavior. If the management of inputs is relatively costly, or ineffective, as it would seem to be in these cases, but, nonetheless if team effort is more productive than separable production with exchange across markets, then there will develop a tendency to use profit-sharing schemes to provide incentives to avoid shirking.<sup>12</sup>

<sup>10</sup> While the degree to which residual claims are centralized will affect the size of the team, this will be only one of many factors that determine team size, so as an approximation, we can treat team size as exogenously determined. Under certain assumptions about the shape of the “typical” utility function, the incentive to avoid shirking with unequal profit-sharing can be measured by the Herfindahl index.

<sup>11</sup> The use of the word *active* will be clarified in our discussion of the corporation, which follows below.

<sup>12</sup> Some sharing contracts, like crop sharing, or rental

### B. Socialist Firms

We have analyzed the classical proprietorship and the profit-sharing firms in the context of free association and choice of economic organization. Such organizations need not be the most viable when political constraints limit the forms of organization that can be chosen. It is one thing to have profit sharing when professional or artistic talents are used by small teams. But if political or tax or subsidy considerations induce profit-sharing techniques when these are not otherwise economically justified, then additional management techniques will be developed to help reduce the degree of shirking.

For example, most, if not all, firms in Yugoslavia are owned by the employees in the restricted sense that all share in the residual. This is true for large firms and for firms which employ nonartistic, or nonprofessional, workers as well. With a decay of political constraints, most of these firms could be expected to rely on paid wages rather than shares in the residual. This rests on our auxiliary assumption that general sharing in the residual results in losses from enhanced shirking by the monitor that exceed the gains from reduced shirking by residual-sharing employees. If this were not so, profit sharing with employees should have occurred more frequently in Western societies where such organizations are neither banned nor preferred politically. Where residual sharing by employees is politically imposed, as in Yugoslavia, we are led to expect that some management technique will arise to reduce the shirking by the central monitor, a technique that will not be found frequently in Western societies since the monitor retains all (or much) of the re-

sidual in the West and profit sharing is largely confined to small, professional-artistic team production situations. We do find in the larger scale residual-sharing firms in Yugoslavia that there are employee committees that can recommend (to the state) the termination of a manager's contract (veto his continuance) with the enterprise. We conjecture that the workers' committee is given the right to recommend the termination of the manager's contract precisely because the general sharing of the residual increases "excessively" the manager's incentive to shirk.<sup>13</sup>

### C. The Corporation

All firms must initially acquire command over some resources. The corporation does so primarily by selling promises of future returns to those who (as creditors or owners) provide financial capital. In some situations resources can be acquired in advance from consumers by promises of future delivery (for example, advance sale of a proposed book). Or where the firm is a few artistic or professional persons, each can "chip in" with time and talent until the sale of services brings in revenues. For the most part, capital can be acquired more cheaply if many (risk-averse) investors contribute small portions to a large investment. The economies of raising large sums of equity capital in this way suggest that modifications in the relationship among corporate inputs are required to cope with the shirking problem

<sup>13</sup> Incidentally, investment activity will be changed. The inability to capitalize the investment value as "take-home" private property *wealth* of the members of the firm means that the benefits of the investment must be taken as annual income by those who are employed at the time of the income. Investment will be confined more to those with shorter life and with higher rates or pay-offs if the alternative of investing is paying out the firm's income to its employees to take home and use as private property. For a development of this proposition, see the papers by Eirik Furobotn and Svetozar Pejovich, and by Pejovich.

payments based on gross sales in retail stores, come close to profit sharing. However, it is gross output sharing rather than profit sharing. We are unable to specify the implications of the difference. We refer the reader to S. N. Cheung.

that arises with profit sharing among large numbers of corporate stockholders. One modification is limited liability, especially for firms that are large relative to a stockholder's wealth. It serves to protect stockholders from large losses no matter how they are caused.

If every stock owner participated in each decision in a corporation, not only would large bureaucratic costs be incurred, but many would shirk the task of becoming well informed on the issue to be decided, since the losses associated with unexpectedly bad decisions will be borne in large part by the many other corporate shareholders. More effective control of corporate activity is achieved for most purposes by transferring decision authority to a smaller group, whose main function is to negotiate with and manage (renegotiate with) the other inputs of the team. The corporate stockholders retain the authority to revise the membership of the management group and over major decisions that affect the structure of the corporation or its dissolution.

As a result a new modification of partnerships is induced—the right to sale of corporate shares without approval of any other stockholders. Any shareholder can remove his wealth from control by those with whom he has differences of opinion. Rather than try to control the decisions of the management, which is harder to do with many stockholders than with only a few, unrestricted salability provides a more acceptable escape to each stockholder from continued policies with which he disagrees.

Indeed, the policing of managerial shirking relies on across-market competition from new groups of would-be managers as well as competition from members within the firm who seek to displace existing management. In addition to competition from outside and inside managers, control is facilitated by the temporary

congealing of share votes into voting blocs owned by one or a few contenders. Proxy battles or stock-purchases concentrate the votes required to displace the existing management or modify managerial policies. But it is more than a change in policy that is sought by the newly formed financial interests, whether of new stockholders or not. It is the capitalization of expected future benefits into stock prices that concentrates on the innovators the wealth gains of their actions if they own large numbers of shares. Without capitalization of future benefits, there would be less incentive to incur the costs required to exert informed decisive influence on the corporation's policies and managing personnel. Temporarily, the structure of ownership is reformed, moving away from diffused ownership into decisive power blocs, and this is a transient resurgence of the classical firm with power again concentrated in those who have title to the residual.

In assessing the significance of stockholders' power it is not the usual diffusion of voting power that is significant but instead the frequency with which voting congeals into decisive changes. Even a one-man owned company may have a long term with just one manager—continuously being approved by the owner. Similarly a dispersed voting power corporation may be also characterized by a long-lived management. The question is the probability of replacement of the management if it behaves in ways not acceptable to a majority of the stockholders. The unrestricted salability of stock and the transfer of proxies enhances the probability of decisive action in the event current stockholders or any outsider believes that management is not doing a good job with the corporation. We are not comparing the corporate responsiveness to that of a single proprietorship; instead, we are indicating features of the corporate structure that are induced by the problem of

delegated authority to manager-monitors.<sup>14</sup>

#### *D. Mutual and Nonprofit Firms*

The benefits obtained by the new management are greater if the stock can be purchased and sold, because this enables *capitalization* of anticipated future im-

provements into present *wealth* of new managers who bought stock and created a larger capital by their management changes. But in nonprofit corporations, colleges, churches, country clubs, mutual savings banks, mutual insurance companies, and "coops," the future consequences of improved management are not

<sup>14</sup> Instead of thinking of shareholders as joint *owners*, we can think of them as investors, like bondholders, except that the stockholders are more optimistic than bondholders about the enterprise prospects. Instead of buying bonds in the corporation, thus enjoying smaller risks, shareholders prefer to invest funds with a greater realizable return if the firm prospers as expected, but with smaller (possibly negative) returns if the firm performs in a manner closer to that expected by the more pessimistic investors. The pessimistic investors, in turn, regard only the bonds as likely to pay off.

If the entrepreneur-organizer is to raise capital on the best terms to him, it is to his advantage, as well as that of prospective investors, to recognize these differences in expectations. The residual claim on earnings enjoyed by shareholders does not serve the function of enhancing their efficiency as monitors in the general situation. The stockholders are "merely" the less risk-averse or the more optimistic member of the group that finances the firm. Being more optimistic than the average and seeing a higher mean value future return, they are willing to pay more for a certificate that allows them to realize gain on their expectations. One method of doing so is to buy claims to the distribution of returns that "they see" while bondholders, who are more pessimistic, purchase a claim to the distribution that they see as more likely to emerge. Stockholders are then comparable to warrant holders. They care not about the voting rights (usually not attached to warrants); they are in the same position in so far as voting rights are concerned as are bondholders. The only difference is in the probability distribution of rewards and the terms on which they can place their bets.

If we treat bondholders, preferred and convertible preferred stockholders, and common stockholders and warrant holders as simply different classes of investors—differing not only in their risk averseness but in their beliefs about the probability distribution of the firm's future earnings, why should stockholders be regarded as "owners" in any sense distinct from the other financial investors? The entrepreneur-organizer, who let us assume is the chief operating officer and sole repository of control of the corporation, does not find his authority residing in common stockholders (except in the case of a take over). Does this type of control make any difference in the way the firm is conducted? Would it make any difference in the kinds of behavior that would be tolerated by competing managers and investors (and we here deliberately refrain from thinking of them as owner-stockholders in the traditional sense)?

Investment old timers recall a significant incidence of nonvoting common stock, now prohibited in corporations whose stock is traded on listed exchanges. (Why prohibited?) The entrepreneur in those days could hold voting shares while investors held nonvoting shares, which in every other respect were identical. Nonvoting share holders were simply investors devoid of ownership connotations. The control and behavior of inside owners in such corporations has never, so far as we have ascertained, been carefully studied. For example, at the simplest level of interest, does the evidence indicate that nonvoting shareholders fared any worse because of not having voting rights? Did owners permit the nonvoting holders the normal return available to voting shareholders? Though evidence is prohibitively expensive to obtain, it is remarkable that voting and nonvoting shares sold for essentially identical prices, even during some proxy battles. However, our casual evidence deserves no more than interest-initiating weight.

One more point. The facade is deceptive. Instead of nonvoting shares, today we have warrants, convertible preferred stocks all of which are solely or partly "equity" claims without voting rights, though they could be converted into voting shares.

In sum, is it the case that the stockholder-investor relationship is one emanating from the *division* of *ownership* among several people, or is it that the collection of investment funds from people of varying anticipations is the underlying factor? If the latter, why should any of them be thought of as the owners in whom voting rights, whatever they may signify or how ever exercisable, should reside in order to enhance efficiency? Why voting rights in any of the outside, participating investors?

Our initial perception of this possibly significant difference in interpretation was precipitated by Henry Manne. A reading of his paper makes it clear that it is hard to understand why an investor who wishes to back and "share" in the consequences of some new business should necessarily have to acquire voting power (i.e., power to change the manager-operator) in order to invest in the venture. In fact, we invest in some ventures in the hope that no other stockholders will be so "foolish" as to try to toss out the incumbent management. We want him to have the power to stay in office, and for the prospect of sharing in his fortunes we buy nonvoting common stock. Our willingness to invest is enhanced by the knowledge that we can act legally via fraud, embezzlement and other laws to help assure that we outside investors will not be "milked" beyond our initial discounted anticipations.

capitalized into present wealth of stockholders. (As if to make more difficult that competition by new would-be monitors, multiple shares of ownership in those enterprises cannot be bought by one person.) One should, therefore, find greater shirking in nonprofit, mutually owned enterprises. (This suggests that nonprofit enterprises are especially appropriate in realms of endeavor where more shirking is desired and where redirected uses of the enterprise in response to market-revealed values is less desired.)

### *E. Partnerships*

Team production in artistic or professional intellectual skills will more likely be by partnerships than other types of team production. This amounts to market-organized team activity and to a non-employer status. Self-monitoring partnerships, therefore, will be used rather than employer-employee contracts, and these organizations will be small to prevent an excessive dilution of efforts through shirking. Also, partnerships are more likely to occur among relatives or long-standing acquaintances, not necessarily because they share a common utility function, but also because each knows better the other's work characteristics and tendencies to shirk.

### *F. Employee Unions*

Employee unions, whatever else they do, perform as monitors for employees. Employers monitor employees and similarly employees monitor an employer's performance. Are correct wages paid on time and in good currency? Usually, this is extremely easy to check. But some forms of employer performance are less easy to meter and are more subject to employer shirking. Fringe benefits often are in non-pecuniary, contingent form; medical, hospital, and accident insurance, and retirement pensions are contingent payments

or performances partly in *kind* by employers to employees. Each employee cannot judge the character of such payments as easily as money wages. Insurance is a contingent payment—what the employee will get upon the contingent event may come as a disappointment. If he could easily determine what other employees had gotten upon such contingent events he could judge more accurately the performance by the employer. He could "trust" the employer not to shirk in such fringe contingent payments, but he would prefer an effective and economic monitor of those payments. We see a specialist monitor—the union employees' agent—hired by them and monitoring those aspects of employer payment most difficult for the employees to monitor. Employees should be willing to employ a specialist monitor to administer such hard-to-detect employer performance, even though their monitor has incentives to use pension and retirement funds not entirely for the benefit of employees.

### *V. Team Spirit and Loyalty*

Every team member would prefer a team in which no one, not even himself, shirked. Then the true marginal costs and values could be equated to achieve more preferred positions. If one could enhance a common interest in nonshirking in the guise of a team loyalty or team spirit, the team would be more efficient. In those sports where team activity is most clearly exemplified, the sense of loyalty and team spirit is most strongly urged. Obviously the team is better, with team spirit and loyalty, because of the reduced shirking—not because of some other feature inherent in loyalty or spirit as such.<sup>15</sup>

<sup>15</sup> *Sports Leagues*: Professional sports contests among teams is typically conducted by a *league* of teams. We assume that sports consumers are interested not only in absolute sporting skill but also in skills *relative* to other teams. Being slightly better than opposing teams enables one to claim a major portion of the receipts; the

Corporations and business firms try to instill a spirit of loyalty. This should not be viewed simply as a device to increase profits by *over-working* or misleading the employees, nor as an adolescent urge for belonging. It promotes a closer approximation to the employees' potentially available true rates of substitution between production and leisure and enables each team member to achieve a more preferred

inferior team does not release resources and reduce costs, since they were expected in the play of contest. Hence, absolute skill is developed beyond the equality of marginal investment in sporting skill with its true social marginal value product. It follows there will be a tendency to overinvest in training athletes and developing teams. "Reverse shirking" arises, as budding players are induced to overpractice hyperactively relative to the social marginal value of their enhanced skills. To prevent overinvestment, the teams seek an agreement with each other to restrict practice, size of teams, and even pay of the team members (which reduces incentives of young people to overinvest in developing skills). Ideally, if all the contestant teams were owned by one owner, overinvestment in sports would be avoided, much as ownership of common fisheries or underground oil or water reserve would prevent overinvestment. This hyperactivity (to suggest the opposite of shirking) is controlled by the league of teams, wherein the league adopts a common set of constraints on each team's behavior. In effect, the teams are no longer really owned by the team owners but are supervised by them, much as the franchisers of some product. They are not full-fledged owners of their business, including the brand name, and can not "do what they wish" as franchises. Comparable to the franchiser, is the league commissioner or conference president, who seeks to restrain hyperactivity, as individual team supervisors compete with each other and cause external diseconomies. Such restraints are usually regarded as anticompetitive, anti-social, collusive-cartel devices to restrain free open competition, and reduce players' salaries. However, the interpretation presented here is premised on an attempt to avoid hyperinvestment in team sports production. Of course, the team operators have an incentive, once the league is formed and restraints are placed on hyperinvestment activity, to go further and obtain the private benefits of monopoly restriction. To what extent overinvestment is replaced by monopoly restriction is not yet determinable; nor have we seen an empirical test of these two competing, but mutually consistent interpretations. (This interpretation of league-sports activity was proposed by Earl Thompson and formulated by Michael Canes.) Again, athletic teams clearly exemplify the specialization of monitoring with captains and coaches; a captain detects shirkers while the coach trains and selects strategies and tactics. Both functions may be centralized in one person.

situation. The difficulty, of course, is to create economically that team spirit and loyalty. It can be preached with an aura of moral code of conduct—a morality with literally the same basis as the ten commandments—to restrict our conduct toward what we would choose if we bore our full costs.

## VI. Kinds of Inputs Owned by the Firm

To this point the discussion has examined why firms, as we have defined them, exist? That is, why is there an owner-employer who is the common party to contracts with other owners of inputs in team activity? The answer to that question should also indicate the kind of the jointly used resources likely to be owned by the central-owner-monitor and the kind likely to be hired from people who are not team-owners. Can we identify characteristics or features of various inputs that lead to their being hired or to their being owned by the firm?

How can residual-claimant, central-employer-owner demonstrate ability to pay the other hired inputs the promised amount in the event of a loss? He can pay in advance or he can commit wealth sufficient to cover negative residuals. The latter will take the form of machines, land, buildings, or raw materials committed to the firm. Commitments of labor-wealth (i.e., human wealth) given the property rights in people, is less feasible. These considerations suggest that residual claimants—owners of the firm—will be investors of resalable capital equipment in the firm. The goods or inputs more likely to be invested, than rented, by the owners of the enterprise, will have higher resale values relative to the initial cost and will have longer expected use in a firm relative to the economic life of the good.

But beyond these factors are those developed above to explain the existence of

the institution known as the firm—the costs of detecting output performance. When a durable resource is used it will have a marginal product and a depreciation. Its use requires payment to cover at least use-induced depreciation; unless that user cost is specifically detectable, payment for it will be demanded in accord with *expected* depreciation. And we can ascertain circumstances for each. An indestructible hammer with a readily detectable marginal product has zero user cost. But suppose the hammer were destructible and that careless (which is easier than careful) use is more abusive and causes greater depreciation of the hammer. Suppose in addition the abuse is easier to detect by observing the way it is used than by observing only the hammer after its use, or by measuring the output scored from a hammer by a laborer. If the hammer were rented and used in the absence of the owner, the depreciation would be greater than if the use were observed by the owner and the user charged in accord with the imposed depreciation. (Careless use is more likely than careful use—if one does not pay for the greater depreciation.) An absentee owner would therefore ask for a higher rental price because of the higher *expected* user cost than if the item were used by the owner. The expectation is higher because of the greater difficulty of observing specific user cost, by inspection of the hammer after use. Renting is therefore in this case more costly than owner use. This is the valid content of the misleading expressions about ownership being more economical than renting—ignoring all other factors that may work in the opposite direction, like tax provision, short-term occupancy and capital risk avoidance.

Better examples are tools of the trade. Watch repairers, engineers, and carpenters tend to own their own tools especially if

they are portable. Trucks are more likely to be employee owned rather than other equally expensive team inputs because it is relatively cheap for the driver to police the care taken in using a truck. Policing the use of trucks by a nondriver owner is more likely to occur for trucks that are not specialized to one driver, like public transit busses.

The factor with which we are concerned here is one related to the costs of monitoring not only the gross product performance of an input but also the abuse or depreciation inflicted on the input in the course of its use. If depreciation or user cost is more cheaply detected when the owner can see its use than by only seeing the input before and after, there is a force toward owner use rather than renting. Resources whose user cost is harder to detect when used by someone else, tend on this count to be owner-used. Absentee ownership, in the lay language, will be less likely. Assume momentarily that labor service cannot be performed in the absence of its owner. The labor owner can more cheaply monitor any abuse of himself than if somehow labor-services could be provided without the labor owner observing its mode of use or knowing what was happening. Also his incentive to abuse himself is increased if he does not own himself.<sup>16</sup>

<sup>16</sup> Professional athletes in baseball, football, and basketball, where athletes having sold their source of service to the team owners upon entering into sports activity, are owned by team owners. Here the team owners must monitor the athletes' physical condition and behavior to protect the team owners' wealth. The athlete has *less* (not, *no*) incentive to protect or enhance his athletic prowess since capital value changes have less impact on his own wealth and more on the team owners. Thus, some athletes sign up for big initial bonuses (representing present capital value of future services). Future salaries are lower by the annuity value of the prepaid "bonus" and hence the athlete has *less* to lose by subsequent abuse of his athletic prowess. Any decline in his subsequent service value would in part be borne by the team owner who owns the players' future service. This does not say these losses of future salaries have no effect on preservation of athletic talent (we are not making a "sunk cost" error). Instead, we assert that the

The similarity between the preceding analysis and the question of absentee landlordism and of sharecropping arrangements is no accident. The same factors which explain the contractual arrangements known as a firm help to explain the incidence of tenancy, labor hiring or sharecropping.<sup>17</sup>

### VII. Firms as a Specialized Market Institution for Collecting, Collating, and Selling Input Information

The firm serves as a highly specialized surrogate market. Any person contemplating a joint-input activity must search and detect the qualities of available joint inputs. He could contact an employment agency, but that agency in a small town would have little advantage over a large firm with many inputs. The employer, by virtue of monitoring many inputs, acquires special superior information about their productive talents. This aids his *directive* (i.e., market hiring) efficiency. He "sells" his information to employee-inputs as he aids them in ascertaining good input combinations for team activity. Those who work as employees or who rent services to him are using him to discern superior combinations of inputs. Not only

---

preservation is reduced, not eliminated, because the amount of loss of wealth suffered is smaller. The athlete will spend less to maintain or enhance his prowess thereafter. The effect of this revised incentive system is evidenced in comparisons of the kinds of attention and care imposed on the athletes at the "expense of the team owner" in the case where athletes' future services are owned by the team owner with that where future labor service values are owned by the athlete himself. Why athletes' future athletic services are owned by the team owners rather than being hired is a question we should be able to answer. One presumption is cartelization and monopsony gains to team owners. Another is exactly the theory being expounded in this paper—costs of monitoring production of athletes; we know not on which to rely.

<sup>17</sup> The analysis used by Cheung in explaining the prevalence of sharecropping and land tenancy arrangements is built squarely on the same factors—the costs of detecting output performance of jointly used inputs in team production and the costs of detecting user costs imposed on the various inputs if owner used or if rented.

does the director-employer "decide" what each input will produce, he also estimates which heterogeneous inputs will work together jointly more efficiently, and he does this in the context of a privately owned market for forming teams. The department store is a firm and is a superior private market. People who shop and work in one town can as well shop and work in a privately owned firm.

This marketing function is obscured in the theoretical literature by the assumption of homogeneous factors. Or it is tacitly left for individuals to do themselves via personal market search, much as if a person had to search without benefit of specialist retailers. Whether or not the firm arose because of this efficient information service, it gives the director-employer more knowledge about the productive talents of the team's inputs, and a basis for superior decisions about efficient or profitable combinations of those heterogeneous resources.

In other words, opportunities for profitable team production by inputs already within the firm may be ascertained more economically and accurately than for resources outside the firm. Superior combinations of inputs can be more economically identified and formed from resources already used in the organization than by obtaining new resources (and knowledge of them) from the outside. Promotion and revision of employee assignments (contracts) will be preferred by a firm to the hiring of new inputs. To the extent that this occurs there is reason to expect the firm to be able to operate as a conglomerate rather than persist in producing a single product. Efficient production with heterogeneous resources is a result not of having *better* resources but in *knowing more accurately* the relative productive performances of those resources. Poorer resources can be paid less in accord with their inferiority; greater accuracy of

knowledge of the potential and actual productive actions of inputs rather than having high productivity resources makes a firm (or an assignment of inputs) profitable.<sup>18</sup>

### VIII. Summary

While ordinary contracts facilitate efficient specialization according to comparative advantage, a special class of contracts among a group of joint inputs to a team production process is commonly used for team production. Instead of multilateral contracts among all the joint inputs' owners, a central common party to a set of bilateral contracts facilitates efficient organization of the joint inputs in team production. The terms of the contracts form the basis of the entity called the firm—especially appropriate for organizing team production processes.

Team productive activity is that in which a union, or joint use, of inputs yields a larger output than the sum of the products of the separately used inputs. This

<sup>18</sup> According to our interpretation, the firm is a specialized surrogate for a market for team use of inputs; it provides superior (i.e., cheaper) collection and collation of knowledge about heterogeneous resources. The greater the set of inputs about which knowledge of performance is being collated within a firm the greater are the present costs of the collation activity. Then, the larger the firm (market) the greater the attenuation of monitor control. To counter this force, the firm will be divisionalized in ways that economize on those costs—just as will the market be specialized. So far as we can ascertain, other theories of the reasons for firms have no such implications.

In Japan, employees by custom work nearly their entire lives with one firm, and the firm agrees to that expectation. Firms will tend to be large and conglomerate to enable a broader scope of input revision. Each firm is, in effect, a small economy engaging in "intra-national and international" trade. Analogously, Americans expect to spend their whole lives in the United States, and the bigger the country, in terms of variety of resources, the easier it is to adjust to changing tastes and circumstances. Japan, with its lifetime employees, should be characterized more by large, conglomerate firms. Presumably, at some size of the firm, specialized knowledge about inputs becomes as expensive to transmit across divisions of the firms as it does across markets to other firms.

team production requires—like all other production processes—an assessment of marginal productivities if efficient production is to be achieved. Nonseparability of the products of several differently owned joint inputs raises the cost of assessing the marginal productivities of those resources or services of each input owner. Monitoring or metering the productivities to match marginal productivities to costs of inputs and thereby to reduce shirking can be achieved more economically (than by across market bilateral negotiations among inputs) in a firm.

The essence of the classical firm is identified here as a contractual structure with: 1) joint input production; 2) several input owners; 3) one party who is common to all the contracts of the joint inputs; 4) who has rights to renegotiate any input's contract independently of contracts with other input owners; 5) who holds the residual claim; and 6) who has the right to sell his central contractual residual status. The central agent is called the firm's owner and the employer. No authoritarian control is involved; the arrangement is simply a contractual structure subject to continuous renegotiation with the central agent. The contractual structure arises as a means of enhancing efficient organization of team production. In particular, the ability to detect shirking among owners of jointly used inputs in team production is enhanced (detection costs are reduced) by this arrangement and the discipline (by revision of contracts) of input owners is made more economic.

Testable implications are suggested by the analysis of different types of organizations—nonprofit, proprietary for profit, unions, cooperatives, partnerships, and by the kinds of inputs that tend to be owned by the firm in contrast to those employed by the firm.

We conclude with a highly conjectural

but possibly significant interpretation. As a consequence of the flow of information to the central party (employer), the firm takes on the characteristic of an efficient market in that information about the productive characteristics of a large set of specific inputs is now more cheaply available. Better recombinations or new uses of resources can be more efficiently ascertained than by the conventional search through the general market. In this sense inputs compete with each other within and via a firm rather than solely across markets as conventionally conceived. Emphasis on interfirm competition obscures intrafirm competition among inputs. Conceiving competition as the *revelation and exchange* of knowledge or information about qualities, potential uses of different inputs in different potential applications indicates that the firm is a device for enhancing competition among sets of input resources as well as a device for more efficiently rewarding the inputs. In contrast to markets and cities which can be viewed as publicly or nonowned market places, the firm can be considered a privately owned market; if so, we could consider the firm and the ordinary market as competing types of markets, competition between private proprietary markets and public or communal markets. Could it be that the market suffers from the defects of com-

munal property rights in organizing and influencing uses of valuable resources?

## REFERENCES

- M. Canes, "A Model of a Sports League," unpublished doctoral dissertation, UCLA 1970.
- S. N. Cheung, *The Theory of Share Tenancy*, Chicago 1969.
- R. H. Coase, "The Nature of the Firm," *Economica*, Nov. 1937, 4, 386-405; reprinted in G. J. Stigler and K. Boulding, eds., *Readings in Price Theory*, Homewood 1952, 331-51.
- E. Furobotn and S. Pejovich, "Property Rights and the Behavior of the Firm in a Socialist State," *Zeitschrift für Nationalökonomie*, 1970, 30, 431-454.
- F. H. Knight, *Risk, Uncertainty and Profit*, New York 1965.
- S. Macaulay, "Non-Contractual Relations in Business: A Preliminary Study," *Amer. Sociological Rev.*, 1968, 28, 55-69.
- H. B. Malmgren, "Information, Expectations and the Theory of the Firm," *Quart J. Econ.*, Aug. 1961, 75, 399-421.
- H. Manne, "Our Two Corporation Systems: Law and Economics," *Virginia Law Rev.*, Mar. 1967, 53, No. 2, 259-84.
- S. Pejovich, "The Firm, Monetary Policy and Property Rights in a Planned Economy," *Western Econ. J.*, Sept. 1969, 7, 193-200.
- M. Silver and R. Auster, "Entrepreneurship, Profit, and the Limits on Firm Size," *J. Bus. Univ. Chicago*, Apr. 1969, 42, 277-81.
- E. A. Thompson, "Nonpecuniary Rewards and the Aggregate Production Function," *Rev. Econ. Statist.*, Nov. 1970, 52, 395-404.

# Education and Underemployment in the Urban Ghetto

By BENNETT HARRISON\*

With the recent availability of several government generated microdata files, it has become possible to develop rigorous quantitative measures of ghetto economic activity, to complement the qualitative impressions contained in the literatures of urban sociology and social psychology.<sup>1</sup> More importantly, analysts are now able to undertake econometric investigations of the structure of ghetto poverty.<sup>2</sup>

The first of these sources, produced by the U.S. Department of Labor, was the 1966 ten ghetto Urban Employment Survey (*UES*). The 37,330 individual person records in the *UES* were assembled from interviews conducted in November 1966. These data facilitated the estimation of ghetto income and unemployment rates, and the comparison of these (and other) statistics with similar indicators for the cities and metropolitan areas of which

these ghettos were a part. In general, as the figures in Table 1 illustrate, the jobs to which ghetto workers have access were found to be of poor quality and paid wages which were substandard by a number of widely accepted benchmarks. Occupation by occupation, the median wage rates of ghetto workers averaged only 40–60 percent of the 1966 annual average wage rates in the corresponding metropolitan area. Given the extent of low-wage work in the slums, it is perhaps not too surprising that so many ghetto men leave (or do not form) families so that mother and children will be eligible for welfare—what amounts to a desperately needed second income. Broken homes in the ghetto may sometimes represent a rational response to the need for multiple incomes. Yet the median incomes of female-headed *UES* households (including welfare receipts), added to the median individual incomes of “unattached” adult males, still sum to less than \$4,000. In 1966, ghetto families with both parents present received only about \$3,500 in gross income. As the data in Table 1 indicate, this is some \$2,500 below the benchmark established by the Department of Labor as a minimum family budget just adequate to sustain an urban family of four in a cheap, rented apartment, with an eight-year-old automobile, and subsisting on a diet consisting largely of dried beans.<sup>3</sup>

<sup>3</sup> See U.S. Bureau of Labor Statistics. For comparative purposes, it should be noted that the income estimates in Table 1 refer to the year preceding November 1966, while the *BLS* budgets refer to March 1967. For an extended discussion of the relevance of the *BLS* cross-section urban budgets as normative benchmarks against which to judge various income distributions, see Vietorisz and Harrison (1971).

\* Associate professor of economics and urban studies, Massachusetts Institute of Technology. This paper is based upon my Pennsylvania Ph.D. thesis. I would like to acknowledge receipt of dissertation fellowships from the Office of Economic Research, Economic Development Administration, U.S. Department of Commerce; and from the Manpower Administration, U.S. Department of Labor. Additional support was provided by the Bureau of Business and Economic Research of the University of Maryland, and by the University's Computer Science Center. Constructive criticism of earlier drafts and materials was contributed by Barbara Bergmann, Norman Glickman, Lawrence Klein, Stephan Michelson, William Milligan, Mancur Olson, Thomas Vietorisz, and several anonymous referees. Marianne Russek provided programming assistance. Elisabeth McDonnell supplied the diagrams. An earlier paper with the same title, reporting on intermediate results of the thesis-in-progress, appeared in David Gordon (1971b).

<sup>1</sup> I refer, for example, to the work of Claude Brown, Kenneth Clark, Elliot Liebow, and Malcolm X.

<sup>2</sup> For some examples, see Peter Doeringer, Gordon, Harrison (1972b), and Thomas Vietorisz and Harrison (1970).

TABLE 1—INCOME, UNEMPLOYMENT, AND SUBEMPLOYMENT IN TEN URBAN GHETTOS

Ghetto and City	Unemployment Rate		Ghetto Subemployment Rate <sup>a</sup>	Median Individual Weekly Wage <sup>a</sup>	Median Annual Family Income <sup>a</sup>	<i>BLS</i>
	Ghetto <sup>a</sup>	SMSA				Lower Level Family Budget <sup>b</sup>
Roxbury (Boston)	6.5	2.9 <sup>a</sup>	24.2	\$74	\$4224	\$6251
Central Harlem (New York City)	8.3	3.7 <sup>a</sup>	28.6	73	3566	6021
East Harlem (New York City)	9.1		33.1	67	3641	
Bedford-Stuyvesant (New York City)	6.3		27.6	73	4736	
North Philadelphia	9.1		34.2	65	3392	5898
North Side (St. Louis)	12.5	4.4 <sup>a</sup>	38.9	66	3544	6002
Slums of San Antonio	7.8	4.2 <sup>b</sup>	47.4	55	2876	n.a.
Mission-Fillmore (San Francisco)	11.4	5.4 <sup>a</sup>	24.6	74	4200	6571
Salt River Bed (Phoenix)	12.5	3.3 <sup>b</sup>	41.7	57	2520	n.a.
Slums of New Orleans	9.5	3.3 <sup>b</sup>	45.3	58	3045	n.a.

Source: Computations from unpublished 1966 *UES* data files and *BLS* worksheets.

<sup>a</sup> November 1966

<sup>b</sup> March 1967

In conducting the 1966 *UES*, the Department of Labor's survey researchers recognized the inability of the conventional measure of unemployment to capture the full force of labor market failure in the ghetto.<sup>4</sup> As the first step toward a remedy, a new index number called the "subemployment rate" was constructed, consisting of the sum of those who are actually unemployed, those working part-time but seeking full-time work, heads of households under 65 years of age earning less than \$60 a week full-time, nonheads under 65 years of age earning less than \$56 a week full-time, half the number of male nonparticipants aged 20-64 (on the

grounds that they have given up looking not because they do not want to work but because of the "conviction—whether right or wrong—that they can't find a job"), and half of the "unfound males."<sup>5</sup>

Among the ten ghettos surveyed by the Department of Labor in 1966 and listed in Table 1, the highest subemployment rate was found in the predominantly Mexican-American slum areas of San Antonio; nearly one out of every two ghetto residents was unemployed or underemployed. The lowest subemployment rate in the sample was 24.2 percent in Boston's Roxbury-South End neighborhoods. Clearly, the problem of underemployment reflected in this indicator lies at the very core of the "urban crisis."

<sup>4</sup> "The traditional unemployment measure counts as working the person who is working only part-time, although he is trying to find full-time work; gives no consideration to the amount of earnings; omits those who are not "actively looking for work"—even though the reason for this is their conviction (whether right or wrong) that they can't find a job, at least one they want; and omits the "undercount" factor—those who are known to be present in the community but who do not show up at all under the present survey methods" U.S. Department of Labor (1967, p. 5).

<sup>5</sup> Highly imperfect and even arbitrary in its weights (if not in its definitions), the index nevertheless represents an extremely important first step toward the measurement of *underemployment* in the United States. Indeed, in all countries—rich and poor alike—the study and the reporting of underemployment as a measure of the *quality* of working life is bound to become increasingly important to the formation of public policy.

From July 1968 through June 1970, the Urban Employment Survey Group of the Department of Labor, together with the Census Bureau, engaged in continuous interviewing of about 3,500 households in each of six central city ghetto areas and two nonghetto "control" areas. The results of these 1969 and 1970 *UES*s are being published gradually by the regional *BLS* offices.

In March 1966, the Census Bureau conducted a Survey of Economic Opportunity (*SEO*) for the Office of Economic Opportunity, with an overall sample size of about 30,000 households (nearly 4,000 adult workers in the sample resided in the central city poverty areas of the twelve largest Standard Metropolitan Statistical Areas (*SMSA*)). About 75 percent of the sample were reinterviewed in March 1967. A large (and growing) number of doctoral dissertations are making use of the *SEO*, the editing of which was performed on contract by the Brookings Institution.

Finally, the *UES* questionnaire was appended to the decennial Census long-form in sixty areas in 1970. When this file is assembled, it will represent by far the largest data source on urban poverty areas ever available.

A wide variety of quantitative studies of the ghetto economy are now possible. One question of considerable theoretical and policy interest is the extent to which the underemployment of ghetto residents is attributable to inadequate human capital formation among the ghetto labor force. This is the particular research issue addressed in the following pages.

### I. Education, Training, and the Urban Ghetto

Neoclassical theories of unemployment and poverty are oriented almost entirely toward the "supply side." Thus, according to Lester Thurow, the originators of the antipoverty program decided that

"poverty was to be eliminated by raising everyone's marginal product to the level where [they] would be able to earn an acceptable income. Education and training programs were to be the principal means for raising marginal products . . . increasing workers' human capital could eliminate poverty . . ." (pp. 91-92).

During the first half of the 1960's, many economists—and most policy makers—supported the contention that public education (and, to a somewhat lesser extent, institutional training) were probably the most effective instruments for combatting poverty.<sup>6</sup>

However, more recent empirical research—much of it using microdata—strongly challenges the conventional wisdom which links nonwhite poverty in particular to inadequate human capital (this literature is surveyed in Harrison (1971b)). The marginal efficiency of nonwhite educational quantity and (in some of the studies) quality has been found to be relatively modest in studies by Ivar Berg, Barbara Bergmann and Gerolyn Lyle, Stephan Michelson (1968), Bradley Schiller, and Randall Weiss. The relative ineffectiveness of various kinds of government training is documented by Berg, Harrison (1972b, chs. 1 and 6), Earl Main, Thomas Ribich, and the U.S. General Accounting Office.

None of these studies has explicitly examined the human capital of *ghetto* residents, however. The poverty areas of our central cities, and the more compact "hardcore" ghetto communities within them, contain families of many races and ethnic origins: Blacks, Puerto Ricans, Chicanos, some American Indians, and substantial numbers of poor whites. There is new evidence that ghetto workers—and Blacks in particular—are investing in themselves through the mechanism of

<sup>6</sup> For a comprehensive review of the literature, see Harrison (1972b), ch. 1.

TABLE 2—EDUCATIONAL ATTAINMENT OF NEGRO AND WHITE GHETTO RESIDENTS  
November 1966

Ghetto	All Persons Age 20 and Over			Persons Age 20-24 Only		
	Sample Size	Median Years Completed <sup>a</sup>	Percent With 12 Or More Years	Sample Size	Median Years Completed <sup>a</sup>	Percent With 12 Or More Years
<b>Negroes</b>						
Roxbury (Boston)	1897	11.5±0.2	45.4	282	12.2±0.3	57.1
Central Harlem (New York City)	2501	11.4±0.2	45.2	214	12.3±0.4	69.2
East Harlem (New York City)	689	10.7±0.4	38.0	85	11.7±0.7	47.1
Bedford-Stuyvesant (New York City)	2711	11.7±0.2	47.3	338	12.3±0.3	64.5
North Philadelphia	2414	10.5±0.2	33.8	271	12.1±1.0	53.5
North Side (St. Louis)	2791	9.7±0.2	29.7	280	12.1±0.4	52.9
Slums of San Antonio	320	11.1±0.6	43.4	31	12.7±0.8	80.6
Mission-Fillmore (San Francisco)	720	12.1±0.3	52.5	113	12.5±0.4	71.7
Salt River Bed (Phoenix)	661	9.0±0.4	23.6	53	11.9±0.9	49.1
Slums of New Orleans	2047	8.8±0.2	23.1	266	11.9±0.5	49.2
<b>Whites<sup>b</sup></b>						
Roxbury (Boston)	858	11.1±0.3	42.8	85	12.3±0.8	61.2
Central Harlem (New York City)	139	10.8±1.0	46.0	17	12.4±2.4	76.5
East Harlem (New York City)	796	10.3±0.4	38.3	91	12.4±0.6	67.0
Bedford-Stuyvesant (New York City)	196	11.4±0.7	46.4	9	11.2±1.6	44.4
North Philadelphia	199	11.7±0.9	48.2	17	12.3±0.9	64.7
North Side (St. Louis)	226	8.8±0.7	27.0	14	12.2±1.4	57.1
Slums of San Antonio	157	11.5±0.8	47.8	15	12.0±2.6	53.3
Mission-Fillmore (San Francisco)	881	12.1±0.3	54.3	123	12.9±0.6	79.7
Salt River Bed (Phoenix)	570	9.2±0.4	28.4	41	12.1±1.0	53.7
Slums of New Orleans	593	9.4±0.4	35.2	55	12.3±1.1	63.6

Source: Computations from unpublished 1966 *UES* data files

<sup>a</sup> 95 percent confidence intervals shown

<sup>b</sup> Excludes Mexican-Americans, Puerto Ricans, and other Spanish-surnamed persons

public education. In fact, Blacks in the urban slums have achieved levels of schooling comparable to those of whites in the same neighborhoods, as shown by the overlapping confidence intervals between the two portions of Table 2, which is constructed from 1966 *UES* materials. Moreover, ghetto Blacks seem to have achieved about the national nonwhite average for years of school completed.<sup>7</sup> In fact, the

gap between the national average school completion rates of young whites and blacks has itself nearly disappeared. By 1970, the interracial difference in median years of schooling for persons under 35 years of age was only .4 (U.S. Department of Commerce, table 65).

Data on institutional training in the ghetto areas of New York City are displayed in Table 3, together with compara-

<sup>7</sup> A third of the nonwhite workers living in the poverty areas of the nation's twelve largest cities in March 1966, had completed at least twelve years of school, according to my calculations from the 1966 *SEO* tapes. About 38 percent of the Blacks living in the ten *UES* areas (identified in Table 2) had attained at least that level of schooling by November 1966. For the nation as a whole in 1966, only about 28 percent of the nonwhites aged 25 or more were high school graduates. While no 1966

national statistics on the educational attainment of younger Blacks have been published, Table 2 shows that, at least in the ghetto, younger Blacks do better than the racial average; they *are* staying in school longer. Nationally, the rate of high school completion for Black adults aged 20 and over is therefore probably somewhat greater than 28 percent. It is the similarity of these national and ghetto estimates which is reported in the text.

TABLE 3—PROPORTION OF LABOR FORCE HAVING COMPLETED  
IN-SCHOOL VOCATIONAL TRAINING OR INSTITUTIONAL  
TRAINING, BY RACE, SEX, AND RESIDENCE  
July 1968–June 1969

(in percent)

	Whites		Blacks	
	Men	Women	Men	Women
Completed in-school vocational training				
New York City ghettos <sup>a</sup>	5	15	11	12
Detroit nonghetto <sup>b</sup>	16	28	16	22
Completed institutional training				
New York City ghettos <sup>a</sup>	22	20	24	17
Detroit nonghetto <sup>b</sup>	41	34	36	20

Source: David Gordon, (1971a) Appendix C, based upon 1969 UES.

<sup>a</sup>  $N=2,651$  individuals aged 16–54 and out of school. Sample areas include Harlem, Bedford-Stuyvesant, and the South Bronx.

<sup>b</sup>  $N=3,899$  individuals aged 16–54 and out of school.

tive figures for nonghetto Detroit. In every case, the incidence of program completion among ghetto dwellers during fiscal year 1969 was lower than for workers residing in the nonghetto control area. *Within* the ghetto, however, Black men compared quite favorably with their white neighbors in terms of the incidence of completion of training programs.

Have ghetto dwellers been able to translate this acquired human capital into improved economic status? Are the inter-racial differences in the marginal efficiency of human capital as significant *within* the ghetto as they are in the nation as a whole? Has ghetto underemployment been relieved by increased education and training? These are the specific questions to which we now turn our attention.

## II. Returns to Human Capital in the Ghetto

### *The Quality of Education*

Before considering my own thesis results on eighteen urban ghettos, a word is in order concerning the devilishly difficult problem of accounting for inter-racial differences in the quality of educa-

tion. To some extent, the discussion has been mooted by Randall Weiss' demonstration of the (rather remarkable) insensitivity of white-nonwhite income differentials to educational "quality" adjustments using the Coleman scores. Moreover, my "controlling" for residence probably includes some adjustment for quality, to the extent that inner city neighbors attend the same (or similar) schools. Finally, it is reasonable to assume that poor white urban adults educated elsewhere (for example, in the rural South or in Appalachia) received relatively low quality schooling. For these reasons, the models employed in my study make no attempt to control for possible variations in educational returns attributable to differences in the quality of education. In this, I am following a precedent set by Gary Becker, who assumed that whites and Negroes of equivalent age and sex were equally productive at the margin.

*Within* racial groups, the implications of not adjusting the data for educational quality are fairly easy to assess. Following Weiss, we may assume that the "quality of schooling, grades, individual ability and

motivation, and *parents'* income, education, and occupation [are] probably . . . correlated positively with years of school [such that] estimates of the increase in earnings associated with an additional year of school will [in the absence of controls for the above factors] be biased upward" (p. 152). It follows that my intraracial estimates of the marginal returns to education are, if anything, too high.<sup>8</sup>

*Human Capital and Underemployment  
in the Central City Poverty Areas of  
Twelve SMSAs<sup>9</sup>*

The wage and employment-determination process which informs these studies of

<sup>8</sup> While we are on the subject of bias in the intraracial regressions, there are two further observations to make, both favorable to the conclusions of this paper. First, Ivar Berg observes that "the basic assumption of the theory of human capital" (i.e., that education and productivity are positively correlated) usually goes untested in most human capital studies—including my own. If earnings are roughly proportional to productivity, then any model which assumes a positive correlation between education and productivity may impart an upward bias to the estimates of the monetary returns to that education. Berg (ch. 5) himself has identified industries in which (he estimates) education and productivity are, if anything, inversely related. Second, Barbara Bergmann believes that the results reported in this paper may be biased upward vis-à-vis appropriate time series estimates. The differences in pay between a (well-educated) Black professional and less-educated Blacks may depend a good deal on the relative scarcity of educated Blacks. Increasing the number of Black professionals by increasing the number of educated Blacks—over time—might reduce the differential between the pay of Blacks with different amounts of education. The result is that a cross-section model predicting income from schooling may exaggerate the effect of improving Black educational achievement on Black income.

<sup>9</sup> From the 61,517 persons aged 14 or more who are described in the March 1966 *SEO*, I selected for analysis those 3,756 persons living in the central city poverty areas of the twelve largest *SMSAs*: Baltimore, Chicago, Cleveland, Detroit, Houston, Los Angeles, New York, Philadelphia, Pittsburgh, St. Louis, San Francisco, and Washington, D.C. These are the only areas which are individually identifiable on the current edition of the *SEO* tapes (while the 1967 file is more recent, only the 1966 file contains data on training program participation). For technical documentation and the definition of "poverty area" as employed in the *SEO*, see J. R. Wetzel and Susan Holland.

In order to permit comparisons among wage rates in

the ghetto labor force is based upon the work of Otis Dudley Duncan and his associates.<sup>10</sup> A fundamentally recursive process is assumed, according to which education, occupational "assignment," and employment (i.e., the practice of an occupation) occur sequentially, with education reentering the system at each "step" because of its convenience to employers as a seemingly inexpensive (although perhaps unreliable) indicator of potential productivities. Wage bargaining, which depends upon the relative economic power of labor and capital, occurs *after* individuals have been "assigned" to occupations. Since economic power varies among industries, the industry attachment of a worker is an important determinant of his or her wage (industry is also a sensible—if crude—proxy for the complementary capital structures with which workers cooperate). The validity of the recursive hypothesis (confirmed by tests on the residuals of the wage and unemployment equations) permits the use of single-equation methods of estimation. Thus, *OLS* was used throughout.

It has been suggested that controlling for industry eliminates one important source of interpersonal wage variation. If serious, this would tend to bias the "pay-

the twelve different cities, it was necessary to construct a cross-section inter-city deflator. The "fixed market bundle" data for such an index was obtained from the *BLS* Lower Level Family Income Schedule for March 1967 (see Bureau of Labor Statistics, bulletin 1570-5; data for the relevant cities are not available for any earlier date). While the size of the sample may appear to be more than adequate, the specification of analytic models using categorical variables rapidly exhausts observations. For individual cities, there are many categories (e.g., "13-15 years of school," "employment in state government," "completion within the last five years of an on-the-job training program") whose cells are empty or nearly so. For this reason, the data from the twelve *SMSAs* have been pooled. Whenever this was done, however, inter-city variations were captured (at least to some extent) by city dummies.

<sup>10</sup> Compare Duncan's methodological exposition and the monograph by Duncan and Peter Blau.

off" coefficients downward. This is, of course, an empirical question. My fifteen industry categories appear to be sufficiently highly aggregated such that there is as much wage and unemployment variation within these major industry groups as there is among them. For example, a model containing education-industry interaction terms was fitted; nearly all of these terms were statistically insignificant. Regressions of earnings and unemployment on age, race, sex, and schooling, using a completely different data source this time, yielded substantially similar estimates of the payoff to education, even though (in this case) industry was not specified (see the following section). Nevertheless, it must be admitted that the inclusion of industry *may* have imparted some degree of downward bias to the results. The aforementioned experimental evidence suggests, however, that it is unlikely to be serious enough to offset the upward biases discussed in footnote 8.

Several different micromodels were specified for each of the two racial groups identifiable in the *SEO*: "whites" (including a small proportion of Spanish-Americans) and "nonwhites" (of whom about 94 percent are Negroes). Multiple regression models with a linear, continuous specification of "education" (measured by years of school completed) indicated that white workers in the twelve central city "ghettos" earn on the average more than twice as much in weekly wages per extra year of school completed as nonwhites from the same neighborhoods. Interaction models showed that the nonwhite payoff varies very much more than the white payoff—from city to city, across the sexes, and by age. Similar contrasts were obtained from models with unemployment specified as the measure of payoff, i.e.,

$$u = \left[ 1 - \frac{\text{weeks worked in 1965}}{\text{weeks in the labor force in 1965}} \right] \times 100$$

The white slope estimate greatly exceeded the nonwhite coefficient in the simple linear analysis, but (again using interaction models) there was considerably more intercity and interage variation about the nonwhite slope than about the white slope, just as before. This same contrast was also found to obtain for two other measures of the returns to schooling: annual wage income of the worker and his or her occupational status.<sup>11</sup>

Perhaps the most interesting of the models developed to study the returns to schooling enjoyed by current ghetto residents were those specifying discontinuous educational steps, designed to capture the discrete effects (if any) of the possession of educational credentials, such as a high school diploma. Let us examine these nonlinear (but additive) models carefully. For each of the two racial groups, four models were estimated:

$$\begin{aligned} (1) \quad \phi &= \beta_0 + \sum_{i=1}^6 \beta_i E_i + \sum_{i=7}^{14} \beta_i T_{i-6} + \beta_{15} S \\ &\quad + \beta_{16} A + \sum_{i=17}^{28} \beta_i C_{i-16} + \epsilon \\ (2) \quad w &= \beta_0 + \sum_{i=1}^6 \beta_i E_i + \sum_{i=7}^{16} \beta_i T_{i-6} + \beta_{17} S \\ &\quad + \beta_{18} A + \sum_{i=19}^{33} \beta_i I_{i-18} \\ &\quad + \sum_{i=34}^{45} \beta_i C_{i-33} + \beta_{46} F + \epsilon \end{aligned}$$

<sup>11</sup> These status calculations employ an ordinal prestige scale which assigns a rank of 0-100 to each of 308 Census occupational titles. This index was developed by the National Opinion Research Corporation (*NORC*) and Otis Dudley Duncan. The properties of the index and a description of the methodology by which it was created are also presented in my dissertation (see Harrison (1972b), Appendix One).

$$(3) \quad Y = \beta_0 + \sum_{i=1}^6 \beta_i E_i + \sum_{i=7}^{16} \beta_i T_{i-6} + \beta_{17} S \\ + \beta_{18} A + \sum_{i=19}^{33} \beta_i I_{i-18} \\ + \sum_{i=34}^{45} \beta_i C_{i-33} + \beta_{46} F + \epsilon$$

$$(4) \quad u = \beta_0 + \sum_{i=1}^6 \beta_i E_i + \sum_{i=7}^{16} \beta_i T_{i-6} + \beta_{17} S \\ + \beta_{18} A + \sum_{i=19}^{33} \beta_i I_{i-18} \\ + \sum_{i=34}^{45} \beta_i C_{i-33} + \beta_{46} F + \epsilon$$

where

$\phi$  = Duncan—NORC ordinal status score  
( $0 \leq \phi \leq 100$ ), based upon occupation  
in March 1966

$w$  = individual weekly earnings in March  
1966

$Y$  = individual annual earnings in 1965

$$u = \left[ 1 - \frac{\text{weeks worked in 1965}}{\text{weeks in the labor force in 1965}} \right] \\ \times 100$$

$$E_1 = E_{0-7 \text{ years}} \\ = \begin{cases} 1 & \text{if individual completed less than 8} \\ & \text{years of school} \\ 0 & \text{otherwise} \end{cases}$$

$$E_2 = E_{8 \text{ years}} \\ = \begin{cases} 1 & \text{if individual completed exactly 8} \\ & \text{years of school} \\ 0 & \text{otherwise} \end{cases}$$

$$E_3 = E_{9-11 \text{ years}} \\ = \begin{cases} 1 & \text{if individual completed some but} \\ & \text{not all high school} \\ 0 & \text{otherwise} \end{cases}$$

$$E_4 = E_{12 \text{ years}} \\ = \begin{cases} 1 & \text{if individual completed high school} \\ 0 & \text{otherwise} \end{cases}$$

$$E_5 = E_{13-16 \text{ years}} \\ = \begin{cases} 1 & \text{if individual completed some but} \\ & \text{not all college} \\ 0 & \text{otherwise} \end{cases}$$

$$E_6 = E_{16^+ \text{ years}} \\ = \begin{cases} 1 & \text{if individual completed at least 4} \\ & \text{years of college} \\ 0 & \text{otherwise} \end{cases}$$

$$T_1 = \begin{cases} 1 & \text{if individual completed a private} \\ & \text{institutional training program since} \\ & \text{1956} \\ 0 & \text{otherwise} \end{cases}$$

$$T_2 = \begin{cases} 1 & \text{if individual started but did not} \\ & \text{complete such a program} \\ 0 & \text{otherwise} \end{cases}$$

$$T_3 = \begin{cases} 1 & \text{if individual completed an appren-} \\ & \text{ticeship since 1956} \\ 0 & \text{otherwise} \end{cases}$$

$$T_4 = \begin{cases} 1 & \text{if individual started but did not} \\ & \text{complete an apprenticeship} \\ 0 & \text{otherwise} \end{cases}$$

$$T_5 = \begin{cases} 1 & \text{if individual completed an on-the-} \\ & \text{job training (OJT) program of at} \\ & \text{least 6 weeks duration since 1956} \\ 0 & \text{otherwise} \end{cases}$$

$$T_6 = \begin{cases} 1 & \text{if individual started but did not} \\ & \text{complete such a program} \\ 0 & \text{otherwise} \end{cases}$$

$$T_7 = \begin{cases} 1 & \text{if individual completed an armed} \\ & \text{forces vocational program since 1956} \\ 0 & \text{otherwise} \end{cases}$$

$$T_8 = \begin{cases} 1 & \text{if individual started but did not} \\ & \text{complete such a program} \\ 0 & \text{otherwise} \end{cases}$$

$$T_9 = \begin{cases} 1 & \text{if individual completed a govern-} \\ & \text{ment training program, e.g., MDTA} \\ & \text{training, Job Corps} \\ 0 & \text{otherwise} \end{cases}$$

$$T_{10} = \begin{cases} 1 & \text{if individual started but did not} \\ & \text{complete such a program} \\ 0 & \text{otherwise} \end{cases}$$

$$S = \begin{cases} 1 & \text{if individual is male} \\ 0 & \text{otherwise} \end{cases}$$

$A$  = individual's age (in years)

$\{I_i\}, i = 1, \dots, 15$

- 1 = construction industry
- 2 = durable goods manufacture
- 3 = nondurable goods manufacture
- 4 = transportation
- 5 = communication
- 6 = utilities, sanitation
- 7 = wholesale trade
- 8 = retail trade
- 9 = finance, insurance, real estate
- 10 = business and repair services
- 11 = personal services
- 12 = entertainment, recreation
- 13 = professional services
- 14 = federal government
- 15 = state or local government

$\{C_i\}, i = 1, \dots, 12$

- 1 = Baltimore
- 2 = Chicago
- 3 = Cleveland
- 4 = Detroit
- 5 = Houston
- 6 = Los Angeles
- 7 = New York
- 8 = Philadelphia
- 9 = Pittsburgh
- 10 = St. Louis
- 11 = San Francisco
- 12 = Washington, D.C.

$$F = \begin{cases} 1 & \text{if individual worked at least 35} \\ & \text{hours/week for at least half of the} \\ & \text{weeks he or she worked in 1965} \\ 0 & \text{otherwise} \end{cases}$$

In all runs,  $(E_1)$ ,  $(I_1)$ , and  $(C_7)$  were deleted to avoid singularity. The absence of the industry and *OJT* dummies and the "full-time/part-time" variable  $F$  from model (1) is motivated by the previously described assumption (following Duncan) that occupational "assignment" precedes actual employment in a specific job in a specific industry.<sup>12</sup>

<sup>12</sup> The variables displayed surprisingly mild multicollinearity. For all but 34 of the 1225 simple correlation coefficients ( $r$ ) associated with the 50 variables ( $x$ ) in the analysis,  $r_{xi x_j} < R_k$ , where  $R_k$  represents the mul-

The estimated education parameters for all eight regressions are displayed in Table 4.<sup>13</sup> For each of the four measures of payoff for each racial group, the table lists the level of the marginal impact of each of the mutually exclusive and exhaustive education classes, and the intraracial differences between successive levels. The weekly wage and annual unemployment *differences* are displayed as step-functions in Figures 1 and 2 (only those steps—i.e., differences—which are significantly different from zero at the .10 level have been graphed). Finally, for each of the payoff variables, Table 4 shows the differences between the white and non-white returns at each level of education. Notice that the great majority of these interracial differences are statistically significant.

Let us first consider the weekly wage indicator. According to Table 4 and Figure 1, the weekly wage of white high school graduates in the pooled set of twelve poverty areas is nearly \$25 higher than that of whites who never entered high school. For nonwhites, the difference is only \$8.33. High school, therefore, has three times as high a marginal payoff for ghetto whites as for ghetto nonwhites.<sup>14</sup> On the assumptions of a forty-year working life, a rectangular lifetime earnings

multiple correlation coefficient of the  $k$ th regression ( $K=1, \dots, 8$ ). (See Donald Farrar and Robert Glauber, p. 98). Even among the 60 human capital variables ( $E_1, \dots, E_6; T_1, \dots, T_{10}$ ) there was little intercorrelation. In fact, only 29 of the 60 simple correlation coefficients  $r_{ET}$  were themselves significant at the .05 level, and in no case was  $r_{ET} \geq R_k$ .

<sup>13</sup> Unfortunately, the training variables in the *SEO* proved to be rather weak. There were too few observations (particularly of persons having completed training) to permit much confidence about the results. Nevertheless, for what it is worth, very few of the training coefficients in any of the regressions were significant at the .05 level.

<sup>14</sup> A 95 percent confidence interval on the interracial difference of \$16.55 is \$2.35–\$30.25.

TABLE 4—MARGINAL RETURNS TO WHITE AND NONWHITE EDUCATION IN THE CENTRAL CITY POVERTY AREAS OF TWELVE SMSA's, March 1966

Measure of Payoff	Years of School Completed					$R^2$	$F$	Sample Size
	8	9-11	12	13-15	16+			
Occupational Status (rank order)								
Whites								
Levels	+2.8	+3.7	+15.8 <sup>a</sup>	+24.1 <sup>a</sup>	+37.9 <sup>a</sup>	.203	16.24	821
Differences		+0.9	+12.1 <sup>b</sup>	+ 8.3 <sup>b</sup>	+13.8 <sup>b</sup>			
Nonwhites								
Levels	+1.7	+2.9 <sup>a</sup>	+10.2 <sup>a</sup>	+17.4 <sup>a</sup>	+40.4 <sup>a</sup>	.309	32.19	2935
Differences		+1.2	+ 7.3 <sup>b</sup>	+ 7.2 <sup>b</sup>	+23.0 <sup>b</sup>			
Absolute difference between white and nonwhite levels	1.1	0.8	5.6 <sup>d</sup>	6.7 <sup>d</sup>	2.5 <sup>e</sup>			
Weekly Wages (dollars)								
Whites								
Levels	+2.47	+14.80	+24.88 <sup>a</sup>	+16.99 <sup>a</sup>	+61.36 <sup>a</sup>	.287	9.68	821
Differences		+12.33 <sup>b</sup>	+10.08 <sup>c</sup>	- 7.89	+44.37 <sup>b</sup>			
Nonwhites								
Levels	+2.33	+1.00	+8.33 <sup>a</sup>	+14.39 <sup>a</sup>	+41.36 <sup>a</sup>	.361	45.80	2935
Differences		-1.33	+7.33 <sup>c</sup>	+ 6.06 <sup>c</sup>	+26.97 <sup>b</sup>			
Absolute difference between white and nonwhite levels	0.14	13.80 <sup>c</sup>	16.55 <sup>c</sup>	2.60	20.00 <sup>b</sup>			
Annual Unemployment (percent)								
Whites								
Levels	+0.4	+0.5	-3.5 <sup>a</sup>	-4.9 <sup>a</sup>	-5.4 <sup>a</sup>	.135	4.36	821
Differences		+0.1	-4.0 <sup>b</sup>	-1.4 <sup>c</sup>	-0.5 <sup>c</sup>			
Nonwhites								
Levels	-0.6	+0.3	-0.4 <sup>a</sup>	-1.9	-1.1	.231	23.54	2935
Differences		+0.9	-0.7 <sup>c</sup>	-1.5	+0.8			
Absolute difference between white and nonwhite levels	1.0	0.8	3.1 <sup>c</sup>	3.0 <sup>c</sup>	4.3 <sup>d</sup>			
Annual Wages (dollars)								
Whites								
Levels	+313.66 <sup>a</sup>	+378.20 <sup>a</sup>	+694.46 <sup>a</sup>	+1016.03 <sup>a</sup>	+3228.02 <sup>a</sup>	.277	8.31	821
Differences		+ 64.54 <sup>c</sup>	+316.26 <sup>b</sup>	+ 321.57 <sup>b</sup>	+2211.99 <sup>b</sup>			
Nonwhites								
Levels	+115.99	+184.24	+300.21 <sup>a</sup>	+ 849.07 <sup>a</sup>	+1514.35 <sup>a</sup>	.425	51.41	2935
Differences		+ 68.25	+117.97 <sup>c</sup>	+ 539.86 <sup>b</sup>	+ 674.28 <sup>b</sup>			
Absolute difference between white and nonwhite levels	147.67	193.96 <sup>c</sup>	394.25 <sup>d</sup>	175.96 <sup>c</sup>	1713.67 <sup>d</sup>			

Source: Calculations from unpublished 1966 SEO data files.

<sup>a</sup> Regression coefficient is significant at the .05 level.

<sup>b</sup> Difference is significant at the .05 level (the standard error of the difference between two coefficients  $\beta_i$  and  $\beta_j$  within a regression is estimated by  $\sqrt{\text{Var}_i + \text{Var}_j - 2 \text{Cov}_{ij}}$ ).

<sup>c</sup> Difference is significant at the .10 level.

<sup>d</sup> Difference is significant at the .05 level (determined by pooling the racial samples, adding education-race interaction terms for each education class, and conducting  $t$ -tests on the coefficients associated with each of these education-race dummies).

<sup>e</sup> Difference is significant at the .10 level.

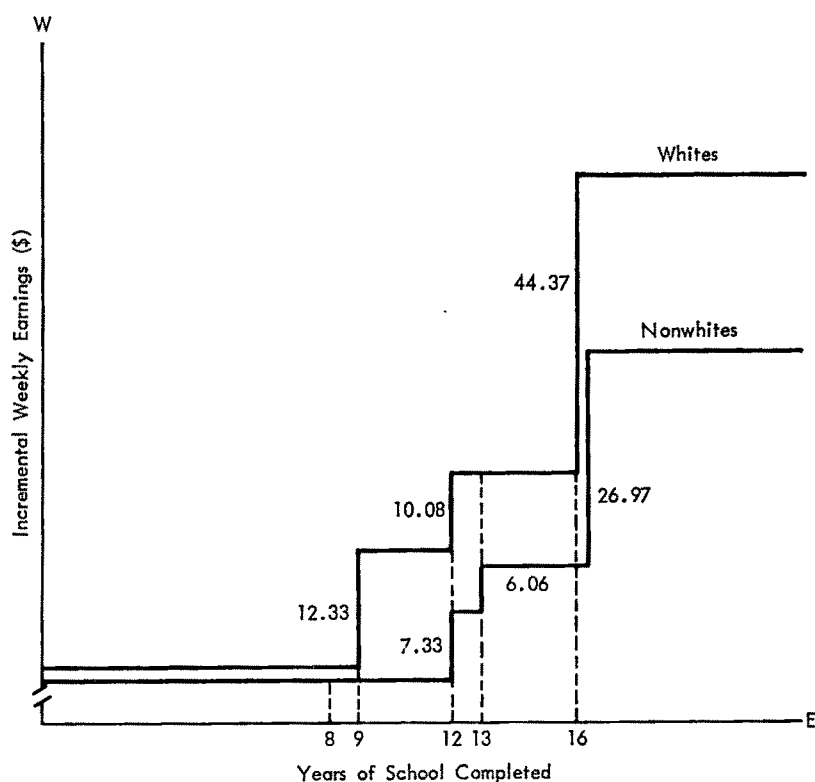


FIGURE 1. INCREMENTAL WEEKLY EARNINGS ASSOCIATED WITH EDUCATION-CENTRAL CITY POVERTY AREAS IN TWELVE SMSAs: MARCH 1966

distribution,<sup>15</sup> and a 6 percent rate of time preference, the present value of the lifetime return to completion of high school is nearly \$19,000 for whites but only \$6,000 for nonwhites. Clearly, education has a very high opportunity cost for nonwhites living in the urban ghetto. There are any number of (largely illegal) activities out "on the street" which are capable of returning at least \$6,000 in a single year.<sup>16</sup>

For whites, the risk of unemployment falls with years of school completed. Over the interval 9-12 years inclusive, the ex-

<sup>15</sup> This is, of course, an unrealistic assumption; age-earnings profiles are known to be hill-shaped. But for the immediate purpose, it is also an inexpensive assumption, and it facilitates computation. Actually, there is some evidence (presented in Harrison (1972b) ch. 2) that a rectangular model gives a tolerably good approximation if one is interested only in the present values computed from an early age.

<sup>16</sup> By assuming equivalent (and continuous) earnings periods for both races, I am surely understating the

expectation of joblessness falls by 4.0 percent (see Figure 2). For white college graduates, it falls an additional 1.9 percent. For nonwhites, on the other hand, the average effect of education on unemployment, while statistically significant, is numerically inconsequential. A white college graduate from the slums can expect to be involuntarily out of work nearly three weeks less per year than a white high school dropout who also lives in one of the urban ghettos in the sample. But the nonwhite college graduate faces only a 0.7 percent lower risk of unemployment than the high school dropout. And that effect is attributable to the high school, not the college, diploma (8.3 per cent of the non-

racial earnings "gap" since nonwhite employment and labor force participation rates are well known to be significantly lower over time than the rates for white workers.

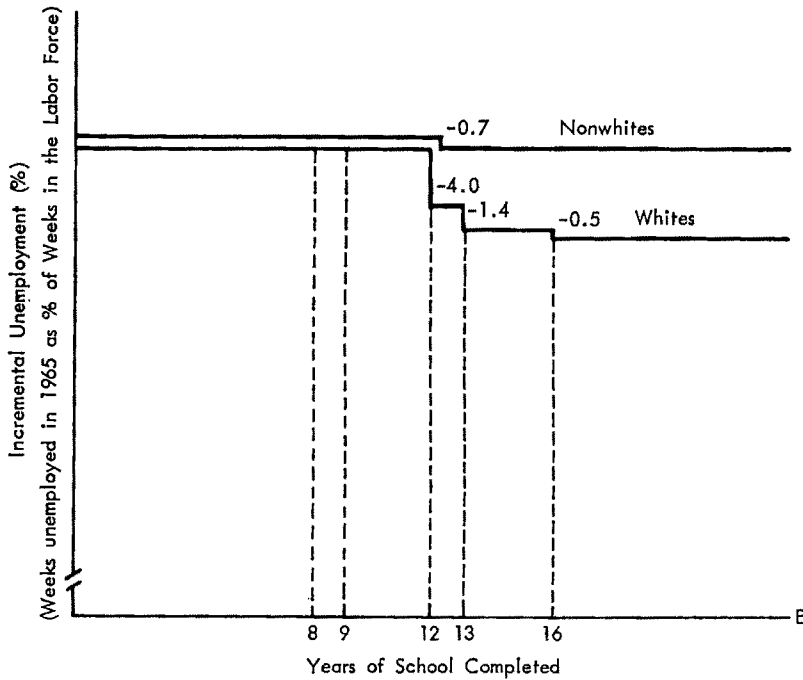


FIGURE 2. INCREMENTAL UNEMPLOYMENT RATES ASSOCIATED WITH EDUCATION—CENTRAL CITY POVERTY AREAS IN TWELVE SMSAs: MARCH 1966

white ghetto sample completed at least some college).

That the annual earnings variable shows the same relative behavior as the other payoff measures, and that education clearly *is* positively associated with occupational status, suggests the following hypothesis: Education facilitates the entry of both white and nonwhite ghetto workers into new occupations, and (at least for whites) leads to greater interoccupational mobility. Moreover, by national standards, these variations represent a "move up" into higher status positions. But this finding only confirms that the effects of racial discrimination pervade even these poorest neighborhoods in the urban economy, for, even though they share many similar problems associated with their common environment, ghetto residents diverge significantly by race insofar as their labor force status is concerned. Education may help members of both races

to move into what are nationally considered to be more prestigious positions. But, once there, the nonwhites find themselves underemployed again: receiving earnings which are hardly above the levels enjoyed in the previous position, and facing the same expectations of unemployment as before. For ghetto whites, on the other hand, the occupational mobility facilitated by education is translated into substantially higher earnings and significantly lower risks of joblessness.

Probably the most dramatic of these findings concerns the virtual absence for nonwhites of any relationship between education and unemployment, after the effects of age, sex, industry, city, training experience, and full-time/part-time status have been removed. In an earlier study of the Harlem economy (Victorisz and Harrison (1970)), it was found that when we stratified both unemployment and labor force participation rates by age, sex, and

years of school completed, the resulting tables displayed a surprising absence of the expected inverse relationship between education and unemployment, or the expected direct relationship between education and labor force participation. In fact, many of the cells showed precisely the opposite effects. From this, we hazarded a tentative explanation, for which I now feel considerably more confidence after completing my thesis research on seventeen additional ghetto areas. Perhaps education increases the expectations and standards of ghetto workers which, when unmet by discriminating or otherwise exploitative employers, leads to frustration. This in turn may reduce the job attachment of the worker. If presently employed, he or she may display greater absenteeism, more frequent recalcitrance when given orders by foremen, or less patience with what is perceived as racist behavior on the part of coworkers.<sup>17</sup> If the ghetto worker is not presently employed, then—although he is indeed searching for work—the change in his standards or expectations may lead him to increase his reservation wage. If the offered positions do not meet his standards, then he will reject the job and search further, or turn to other in-

come-generating activities such as public welfare or “the hustle.”<sup>18</sup> In this way, he may remain unemployed for a relatively long period of time.

*Education and Underemployment in  
Ten Ghettos in Eight SMSAs<sup>19</sup>*

The earnings variable used with this sample is hourly, rather than weekly or annual earnings, and the unemployment variable (a 0–1 dummy) refers to the survey week rather than, as with the earlier sample, to the previous year. In the ten urban ghettos studied, education has only a limited impact upon hourly earnings, and virtually no effect at all upon what must now be interpreted as the conditional probability of unemployment in any given week (since this sample consisted almost entirely of Black, Puerto Rican, and Chicano workers, white/non-white differences were not studied at all).

A number of different earnings models yielded estimated returns of from 3 to 9 cents per hour for each additional year of schooling and, over the interval of 9–12 years of schooling inclusive, an average return of 15 cents. Workers with at least some college received, on the average, only 20 cents more per hour than high school graduates who went directly to work and did not go to college. One model was designed to test for the existence of an upper limit to hourly wages and salaries in the ghetto, irrespective of education, i.e., the upper asymptote  $e^a$  in the function

$$w = e^{a-bE-1}$$

<sup>17</sup> Harold Sheppard's famous studies of ‘workers with the ‘blues’ ’ have shown that “the greater a person's education achievements, as measured by years of school, the greater are his life and job aspirations,” and the greater his discontent if he fails to achieve those goals (1971, p. 28). In their study of Boston labor market institutions, Doeringer and his colleagues found that ghetto job-seekers were systematically placed in jobs similar (in wage and working conditions) to those which they had just left. This lack of upward mobility in the placement system probably contributes to the poor work habits (such as tardiness and high quit rates) which then frighten off other and perhaps genuinely concerned businessmen. Doeringer hypothesizes that “the availability of alternative low wage job opportunities and the unattractiveness of such low wage work interact to discourage the formation of strong ties to particular employers . . . for wage rates higher than the ‘prevailing’ ghetto wage, disadvantaged workers are more likely to be stable employees than other workers” (pp. 10–11).

<sup>18</sup> The functional similarities and complementarities between these “irregular” activities and the low-wage jobs which make up what some economists call the “secondary labor market” are explored in Daniel Fusfeld, Harrison (1972b, ch. 5), and U.S. Department of Labor (1968, pp. 94–99).

<sup>19</sup> The *US* areas in this sample are identified in Table 1. Usable records on 37,330 individuals aged 14 and over were available. The wage data were again adjusted to control for intercity variations in the cost of living.

Such limits were indeed found, ranging from a low of \$1.56 per hour in San Antonio's Chicano *barrios* to a high of \$2.04 per hour in the Bedford-Stuyvesant ghetto of New York City. No amount of education would lead us to expect a Bedford-Stuyvesant worker of having an hourly wage of more than \$2.04. All ten asymptotes were statistically significant. Interestingly, Doeringer had identified the same threshold in his own studies of urban poverty in Boston. While Boston's antipoverty agency set up neighborhood employment centers in the slums to recruit ghetto workers for Boston industry, "many of these jobs pay only \$2.00 an hour or less" (p. 9). Thus, says Doeringer, "the main benefits of the system have come from prompt referrals to jobs similar to those already available to the ghetto community" (p. 17).

### III. Minority Economic Opportunity Outside the Ghetto<sup>20</sup>

It has been suggested that a sample of ghetto residents is inherently biased, since those for whom education has "paid off" will presumably have moved out, leaving us with a sample skewed toward the "failures." A very recent (and decidedly unofficial) finding of the Bureau of Labor Statistics' Urban Employment Survey Group provides us with some direct evidence that the ghetto samples are probably not biased, at least not because of selective outmigration. Of 7,200 ghetto families in six cities who were to be reinterviewed by the BLS over a twelve-month period in 1968-69 (a period of exceptionally high mobility nationally), only 900 had moved from one residence to another. And of these 900 families, only 60 had moved out of the ghetto; all of the rest were either intraghetto moves (750) or involuntary relocations, for example, to

jail or into the armed forces (90). In other words, the rate of family outmigration from the urban ghetto, only two years after the date of the surveys we are analyzing, was less than 1 percent. According to Anthony Downs, p. 36, "normal population turnover causes about 20 percent of the residents of the average U.S. neighborhood to move out each year."<sup>21</sup>

This same question about selective outmigration motivated the extension of my researches to urban workers living outside of the ghettos, (a) in nonpoverty central city neighborhoods, and (b) in suburban communities (the source of this intrametropolitan data is, again, the 1966 *SEO*). The results of these studies seem to validate the earlier findings. More important, they cast considerable doubt on the validity of widely held assumptions about the intrametropolitan spatial distribution of poverty and discrimination.

In terms of average economic opportunity—measured by weekly earnings, annual unemployment, and occupational status—the white levels improve monotonically with "distance" from the core, while nonwhite opportunity increases somewhat with "move" from the ghetto to the nonpoverty central city, but falls again with the further move out to the suburban ring.<sup>22</sup> (See Table 5 where only the male figures are tabulated.) When 95 percent confidence intervals are constructed around these means or medians,

<sup>21</sup> In 1969-70, 25 percent of all Black households in America changed residence (U.S. Department of Commerce, Table 9). In the *UES* sample, as indicated in the text, only 900 of 7,200 households (13 percent) changed residences in 1968-69. It may be worth noting that, unencumbered by the obstacle of purely racial discrimination, the "leakage" of highly motivated and able ghetto whites is at least as likely as the loss of similarly equipped Blacks. If so, then the results reported in this paper imply that even the *least* able ghetto whites receive a significantly higher return to schooling than do ghetto Blacks.

<sup>22</sup> The reader is to be reminded that we are in fact dealing with a 1965 cross-section.

<sup>20</sup> For a more extensive "inside-outside" comparison, using *SEO* data, see Harrison (1972a), (1972b, ch. 4).

TABLE 5—INDICATORS OF RELATIVE ECONOMIC OPPORTUNITY IN TWELVE SMSAs, BY RACE AND LOCATION OF RESIDENCE

March 1966  
(95% confidence intervals shown)

Residential location	Sample Size	Median weekly male earnings	Mean male unemployment rates	Median male occupational status <sup>a</sup>	Incremental return to high school (both sexes) <sup>b</sup>	Incremental return to college (both sexes) <sup>c</sup>
Central city poverty areas						
whites	821	\$93.33 ± 5.76	8.8% ± 1.8	19.4 ± 2.0	\$24.88	\$36.48
nonwhites	2935	78.19 ± 2.36	10.4 ± 1.2	14.7 ± 0.9	8.83	33.03
Rest of central city						
whites	2125	123.67 ± 4.88	3.9 ± 0.8	36.8 ± 1.4	9.16	51.62
nonwhites	1565	99.87 ± 4.16	5.3 ± 1.2	16.7 ± 1.6	<sup>d</sup>	53.50
Suburban ring						
whites	3163	133.58 ± 4.68	3.5 ± 0.6	40.7 ± 1.1	32.80	65.32
nonwhites	845	96.12 ± 5.32	8.8 ± 1.9	15.7 ± 1.8	<sup>d</sup>	38.87

Source: Calculations from unpublished 1966 SEO data files.

<sup>a</sup> Measured by a scoring procedure which assigns an ordinal rank of 0-100 to each of the 308 occupational titles on the SEO tapes.

<sup>b</sup> Measured relative to a person who never attended high school at all. Each of the component "steps" is statistically significant at the .05 level. Estimated by multiple regression, controlling for age, sex, industry, city of residence, and training program experience.

<sup>c</sup> Measured relative to a high school graduate who never attended college at all.

<sup>d</sup> Statistically insignificant at the .05 level.

the contrast is even more dramatic. For whites, employment opportunity definitely rises (or at least does not fall) as we move from the innermost to the outermost sample areas. For nonwhites, however, the three descriptors of employment opportunity show relatively little sensitivity to intrametropolitan residential location. Nonwhite earnings are significantly higher outside the ghetto than inside, but—once "outside"—there is no significant difference between the median levels associated with central city as against suburban residence. Nonwhite unemployment rates in the ghetto and in the suburbs are not statistically different from one another, and may be only slightly lower in the nonpoverty central city. Finally, the indicator of occupational status for nonwhite men is totally insensitive to residential location.

Nor are the marginal weekly wage returns to nonwhite education significantly

greater in the suburbs than in the ghetto. Indeed, the high school diploma as a terminal credential is less valuable to an outside nonwhite than to a nonwhite residing within the ghetto. This controversial empirical result was predicted several years ago by Stephan Michelson (1969), in a behavioral model suggesting that, unless a Black high school student is certain that he will complete college, he may be better off (in terms of his lifetime income) by dropping out of the educational system before finishing high school.

#### IV. Conclusions

Do these findings suggest that public investment in ghetto schools should be cut back? I believe not. In Michelson's words: "Equal education should be a goal in itself, not diminished for its failure to produce income for nonwhites" (1968, pp. 8-28). Continued education should be an important objective of all those concerned

with improvement in the lives of ghetto residents. Certainly any final judgment should be withheld pending the availability of true longitudinal data (indeed, Finis Welch has discovered evidence that Black-white income differences may be shrinking over time, although not necessarily in the ghetto).

Nevertheless, as a short-run antipoverty policy instrument, education without a supply of jobs which utilize and reward the capabilities of ghetto workers is unlikely to have much impact. The prevalence of ghetto unemployment, involuntary part-time employment, and substandard wages, in conjunction with these new findings on nonwhite educational attainment and the recent studies of the relatively modest technical skills required for the "average" performance of an extremely broad range of "typically urban" jobs,<sup>23</sup> strongly suggests that existing urban labor markets under-utilize ghetto workers and do not permit these individuals to realize their potential productivities. If this interpretation is correct, then the remedy must be sought in opening up *new* urban job markets to the ghetto poor, markets whose jobs are physically accessible to ghetto residents, whose availability is made known to them, and whose entry level wages and promotional possibilities will in fact lead to a significant improvement in their levels of living.<sup>24</sup>

<sup>23</sup> Compare Berg, Charlotte Fremon, and Harrison (1972b) ch. 1.

<sup>24</sup> Federally supported and widely publicized attempts to open up *existing* urban labor markets to minorities have not been particularly successful. In the construction industry, for example, "... the Philadelphia Plan, which would compel the hiring of members of minority groups on federal projects totaling \$500,000 or more ... has not begun to produce even minimal gains toward its modest goal of breaking the color barrier in six construction trades. As a result, the Department of Labor is moving to sue a number of contractors" (New York Times, p. 1). The only other significant minority hiring effort currently underway is the Job Opportunities in the Business Sector (JOBS) program being implemented by the National Alliance of Businessmen with Labor Department subsidies. Re-

In other words, the findings reported in this paper seem to call rather convincingly for a change in emphasis away from concentration on the alleged defects of the ghetto poor themselves toward the investigation of defects in the market system which constrains the poor from realizing their potential. Without a direct transformation and augmentation of the demand for their labor, significant improvement in the economic situation of ghetto dwellers is unlikely.<sup>25</sup> Attempts to change the worker himself—whether to remedy his personal "defects" or to move him to a "better" environment—have not worked up until now, and the several new micro-data sources explored in this study provide little if any evidence to support the belief that such attempts will be sufficient in the future.

#### REFERENCES

- G. Becker, *The Economics of Discrimination*, Chicago 1957.
- I. Berg, *Education and Jobs: The Great Training Robbery*, New York 1970.
- B. R. Bergmann and G. Lyle, "The Occupational Standing of Negroes by Areas and Industries," *J. Hum. Resources*, fall 1971, 6, 411-33.
- P. Blau and O. D. Duncan, *The American Occupational Structure*, New York 1967.
- P. B. Doeringer, "Ghetto Labor Markets—Problems and Programs," Program on Regional and Urban Economics, Disc. Paper No. 35, Harvard Univ., 1968.
- A. Downs, "Alternative Futures for the American Ghetto," in A. Downs, ed., *Urban Prob-*

cently completed studies by the Senate Subcommittee on Employment, Manpower and Poverty and by the U.S. General Accounting Office suggest that "many (JOBS) employers are ... taking federal money for hiring people in low-paying jobs that require little or no training" and that, in any case, "many of the larger firms ... have had to lay off many of the trainees they took on in 1968 and 1969" (Washington Evening Star, p. A-20). These programs are discussed at length in Harrison (1972b), (ch. 1, 6) and the U.S. Senate report.

<sup>25</sup> This conclusion constitutes one of the strongest arguments for a program of public service jobs for the disadvantaged. See Harrison (1971a) and Sheppard (1969).

- lems and Prospects*, Chicago 1970.
- O. D. Duncan, "A Socioeconomic Index for all Occupations," in A. J. Reiss, Jr., ed., *Occupations and Social Status*, Glencoe 1951.
- D. E. Farrar and R. R. Glauber, "Multicollinearity in Regression Analysis: The Problem Revisited," *Rev. Econ. Statist.*, Feb. 1967, 49, 92-107.
- C. Fremon, "The Occupational Patterns in Urban Employment Change, 1965-67," Urban Institute Paper No. 113-32, Washington, Jan. 1970.
- D. Fusfeld, "The Basic Economics of the Urban and Racial Crisis," in *Conference Papers of the Union for Radical Political Economics*, Ann Arbor 1968; reprinted in the *Rev. Black Polit. Econ.*, spring-summer 1970, 1, 58-83.
- D. M. Gordon, (1971a) "Class, Productivity, and the Ghetto," unpublished doctoral dissertation, Harvard Univ. 1971.
- , (1971b), *Problems in Political Economy: An Urban Perspective*, Lexington 1971.
- B. Harrison, (1971a), "Public Employment and Urban Poverty," Urban Institute Paper No. 113-43, Washington, June 1971.
- , (1971b), "Human Capital, Elack Poverty, and 'Radical' Economics," *Ind. Rel.*, Oct. 1971, 10, 277-86.
- , (1972a), "The Intrametropolitan Distribution of Minority Economic Welfare," *J. Reg. Sci.*, Apr. 1972, 12, 23-43.
- , (1972b), *Education, Training, and the Urban Ghetto*, Baltimore 1972.
- E. Main, "A Nationwide Evaluation of M.D.T.A. Institutional Job Training," *J. Hum. Resources*, spring 1968, 3, 159-70.
- S. Michelson, "Incomes of Racial Minorities," unpublished doctoral dissertation, Stanford Univ. 1968.
- , "Rational Income Decisions of Negroes and Everybody Else," *Ind. Labor Relat. Rev.*, Oct. 1969, 23, 15-28.
- T. Ribich, *Education and Poverty*, Washington 1968.
- B. Schiller, "Class Discrimination Versus Racial Discrimination," *Rev. Econ. Statist.*, Aug. 1971, 53, 263-69.
- H. L. Sheppard, *The Nature of the Job Problem and the Role of New Public Service Employment*, Kalamazoo 1969.
- , "Discontented Blue-Collar Workers," *Mon. Labor Rev.*, Apr. 1971, 94, 25-32.
- L. Thurow, "Raising Incomes Through Manpower Training Programs," in A. H. Pascal, ed., *Contributions to the Analysis of Urban Problems*, Santa Monica 1968.
- T. Vietorisz and B. Harrison, *The Economic Development of Harlem*, New York 1970.
- and ———, "Ghetto Development, Community Corporations, and Public Policy," *Rev. Black Polit. Econ.*, fall 1971, 2, 28-43.
- R. Weiss, "The Effect of Education on the Earnings of Blacks and Whites," *Rev. Econ. Statist.*, May 1970, 52, 150-159.
- F. Welch, "Black-White Differences in Returns to Schooling," paper delivered to the Princeton Conference on the Economics of Discrimination, mimeo. Sept. 1971.
- J. R. Wetzel and S. B. Holland, "Poverty Areas of Our Major Cities," *Mon. Labor Rev.*, Oct. 1966, 89, 1105-10.
- New York Times*, July 20, 1970.
- U.S. Bureau of the Census, *Current Population Reports*, Series P-23, No. 38, "The Social and Economic Status of Negroes in the United States, 1970," Washington 1971.
- U.S. Department of Labor, "A Sharper Look at Unemployment in U.S. Cities and Slums," Washington 1967.
- , *Manpower Report of the President: 1968*, Washington 1968.
- U.S. Bureau of Labor Statistics, "1966 Urban Employment Survey," computer tapes and mimeographed worksheets.
- , "Three Standards of Living for an Urban Family of Four Persons: Spring, 1967," Bull. 1570-5, Washington 1967.
- U.S. General Accounting Office, "Improvements Needed in Contracting for On-the-Job Training Under the Manpower Development and Training Act of 1962," Washington 1968.
- U.S. Office of Economic Opportunity, "1966 Survey of Economic Opportunity," computer tapes and mimeographed codebooks.
- U.S. Senate, Committee on Labor and Public Welfare, *The J.O.B.S. Program*, Washington 1970.
- Washington Evening Star*, Apr. 29, 1970.

# Option Demand and Consumer's Surplus: Valuing Price Changes under Uncertainty

By RICHARD SCHMALENSEE\*

The concerns of this paper are best introduced through an example. Suppose that my tastes and income next summer are uncertain. I will have a strong desire to visit Yellowstone Park with probability .3, while with probability .7 I will have no wish to go. If the first state of nature materializes, I will be willing to pay up to \$100 above the cost of the trip at the prices which I expect to prevail next summer to guarantee my ability to see Old Faithful, while I would pay nothing in the second state. What is the maximum amount I would pay today (ignoring discounting) to ensure my opportunity to purchase a trip to Yellowstone next summer at the prices I now expect to encounter then?

Burton Weisbrod coined the term "option value" to describe this magnitude, the maximum amount a consumer with uncertain future preferences (and possibly income) would be willing to pay for the option to purchase a particular commodity at a specified price. Millard Long argued that option value was nothing more than the expected value of the consumer's surplus associated with the commodity in question, \$30 in our example, while Cotton Lindsay disagreed. Most recently, John Krutilla and his associates argued that option value generally exceeds expected

consumer's surplus, so that I should be willing to pay more than \$30 to ensure my ability to travel to Yellowstone next summer.<sup>1</sup>

Using the terminology of Krutilla et al., which differs from that of the other papers cited but is somewhat clearer, the general situation we are concerned with can be formalized as follows. Consider an individual in a timeless world, or one with a riskless interest rate of zero, who is uncertain about his preferences and/or his income. Let there be  $N$  possible states of nature, and let the known probability of state  $i$  occurring be  $\Pi_i$ . Consider two price systems,  $P$  and  $P^*$ , where  $P^*$  is the preferred system. In the usual discussions of this problem, and in the example with which we began, all prices in  $P$  and  $P^*$  are assumed the same except for one. That price is low or at least finite in  $P^*$  (the good is available) and very high or infinite in  $P$  (the good is not available).

Let  $S_i$  be the consumer's surplus generated in state  $i$  if the individual is allowed to trade at prices  $P^*$  instead of  $P$ .<sup>2</sup> We let the Option Price ( $OP$ ) be the maximum amount the individual would be willing to pay with certainty to guarantee that price system  $P^*$  would prevail in all states of nature rather than  $P$ . Option Value ( $OV$ ) is then defined as the difference between option price and expected consumer's surplus:

\* Assistant professor of economics, University of California, San Diego. I would like to thank William Brainard, Richard Emmerson, Wolfhard Ramm, and Larry Ruff for many helpful suggestions and discussions and anonymous referees for useful comments on earlier versions of this paper. Errors are, of course, entirely my responsibility.

<sup>1</sup> See also Charles Cicchetti and A. Myrick Freeman III.

<sup>2</sup> Precise definitions of "consumer's surplus", "option price", and "option value" are given in Section II.

$$(1) \quad OV = OP - \sum_{i=1}^N \Pi_i S_i$$

Krutilla et al. have asserted, and Weisbrod and Lindsay have argued implicitly, that  $OV$  is generally positive. Since  $OP$  is the most compelling measure of the value of a shift from  $P$  to  $P^*$ , this contention implies that expected consumer's surplus understates the benefits of price changes when there is uncertainty.

The next section presents the basic state-preference approach employed in this essay and obtains a useful theorem about the nature of risk aversion.<sup>3</sup> Section II proves that if certain markets do not exist (which generally do not exist in real economies), option value depends on the details of individual preferences and circumstances and may be either positive or negative.<sup>4</sup> We then consider situations in which the above-mentioned markets do exist, and we discuss conditions under which option value is zero. The final section briefly examines the implications of our analysis for government investment decisions.<sup>5</sup>

<sup>3</sup> Basic references for the state-preference approach are Kenneth Arrow (1964), Gerhard Debreu, ch. 7, and Jack Hirschleifer. For an alternative analysis of uncertain preferences, see Richard Zeckhauser and the discussion by David Cass.

<sup>4</sup> D. R. Byerlee has recently argued this under very restrictive assumptions.

<sup>5</sup> Throughout this paper we consider only completely irreversible decisions. Either  $P$  or  $P^*$  will prevail in the entire (one period) future; no matter what state of nature occurs, this cannot be altered. More complicated situations involve possible future reconsideration of decisions made today, with incomplete reversibility.

Consider, for instance, a two-period model. Suppose  $P^*$  is purchased by society at a cost of  $C$ . At the end of one period, some new information will be obtained about the future. Depending on the content of this new knowledge,  $P^*$  can be allowed to prevail in the second period, or  $P$  can be reestablished at a cost of  $C'$ . Similarly, if it is decided not to purchase  $P^*$  today, it can be obtained for the second period at a cost of  $C$ . The larger is  $C'$ , the cost of reversing the decision to purchase  $P^*$ , the less attractive is the immediate purchase of that price system. Further analysis of situations of this sort is, however, well beyond the scope of this paper.

## I. A Basic Result

Consider an individual who knows the true probabilities, the  $\Pi_i$ , that each of the  $N$  possible states of nature will occur. Let  $X_i$  be the bundle of commodities the individual will consume if state  $i$  occurs. Using the expected utility theorem of John von Neumann and Oscar Morgenstern and an extension of it by Hirschleifer, the individual's utility function may be represented as

$$(2) \quad V = \sum_{i=1}^N \Pi_i V^i(X_i)$$

where the conditional utility functions  $V^i$  may be different. The individual is assumed to maximize  $V$  subject to whatever budgetary and other constraints are imposed on him.

We assume that the individual knows the set of prices,  $P_i$ , that will prevail the  $i$ th state ( $i=1, \dots, N$ ). Then, as Arrow (1964) has shown, if the individual will receive a known contingent income  $Y_i$  if state  $i$  occurs ( $i=1, \dots, N$ ), we may replace (2) by

$$(3) \quad V = \sum_{i=1}^N \Pi_i U^i(Y_i, P_i)$$

The indirect conditional utility functions in (3) correspond in every state of nature to the conditional utility functions in (2). We shall work only with (3) in the remainder of this paper.

Further, in the remainder of this paper the price sets ( $P_i$ ) are, for simplicity, assumed identical for all states. Given a price system, the conditional income claims ( $Y_i$ ) are the objects of choice which determine the level of utility attained.

Following the literature, we assume that  $V$  is a quasi-concave function of the  $Y_i$ .<sup>6</sup>

<sup>6</sup> See Arrow (1964) and Hirschleifer for discussions of this assumption. Quasiconcavity implies that the set of vectors  $Y$ , with elements  $Y_1, \dots, Y_N$ , for which  $V(Y) \geq V(Y_0)$  is convex for any vector  $Y_0$ .

Arrow and Alain Enthoven have shown that a sufficient condition for (3) to be quasiconcave in the  $Y_i$  is that each of the  $U^i$  be concave functions of the corresponding  $Y_i$ .<sup>7</sup> We assume that this condition holds, and we further assume throughout that  $\partial U^i(Y_i, P_i)/\partial Y_i \equiv U_y^i(Y_i, P_i)$  exists and is differentiable for all  $i$  and for all  $Y_i$  and  $P_i$ .

Sections II and III are concerned with an individual's reaction to changes in the  $Y_i$ . Since the  $U^i$  may each be different functions, though, it is first necessary to find some sensible way to make them comparable. The necessary tool is developed in the following discussion of risk aversion.<sup>8</sup>

Consider an individual who will face the same price system,  $P$ , in all states of nature. Let his conditional income in state  $i$  be  $Y_i$  ( $i=1, \dots, N$ ), as before, and denote the set of these conditional incomes by  $[Y_i]$ . Consider a gamble which will pay this individual  $w_i$  if the  $i$ th state occurs. Recall the usual definition of risk aversion: the individual is said to be risk averse when prices are  $P$  and his conditional incomes are  $[Y_i]$  if, and only if, all gambles which he would accept at this point (denoted hereafter by  $([Y_i], P)$ ) have the property

$$(4) \quad \sum_{i=1}^N \Pi_i w_i > 0$$

It follows that a risk-averse individual does not accept fair or unfair gambles (gambles with zero or negative expected value, respectively); any he does accept must be biased in his favor.

It is easy to show that concavity of all the  $U^i$  is not sufficient to guarantee risk aversion at all  $([Y_i], P)$  points. Suppose the  $U^i$  are strictly concave and identical, but the  $Y_i$  differ. If  $Y_j$  exceeds  $Y_k$  and the

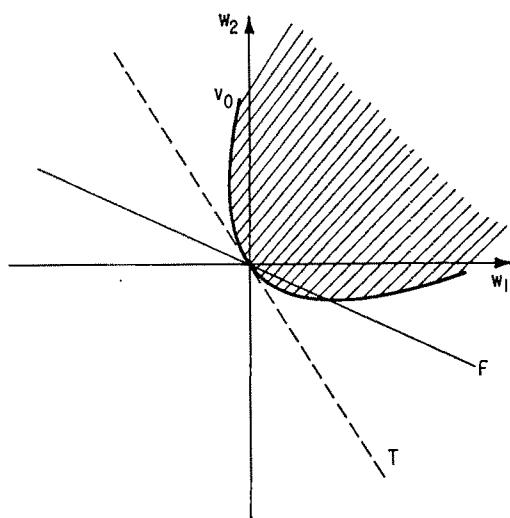


FIGURE 1

two states are equally likely to occur, the individual will accept a fair gamble which adds  $(Y_j - Y_k)/2$  to his income in state  $k$  and subtracts the same amount from his income in state  $j$ . If all states are equally likely, it is obvious that the individual will be risk averse only at points where all the  $Y_i$  are equal.<sup>9</sup>

We now return to the general case, where the  $U^i$  need not be identical. Consider Figure 1, which depicts situations in which only  $w_1$  and  $w_2$  are nonzero.  $V_0$  is locus of points which have the same expected utility, the same value of  $V$ , as the no-gamble situation  $w_1 = w_2 = 0$ . The convex set of points to the northeast of  $V_0$  represents the set of all gambles that would be accepted by this individual at the  $([Y_i], P)$  point considered. The line  $F$  is the locus of all fair gambles; all points to the southwest of this line correspond to unfair gambles. It is clear from the figure that some unfair gambles would be ac-

<sup>7</sup> The necessary and sufficient condition (see Arrow and Enthoven, pp. 799-800) is that at most one of the  $U^i$  be convex in  $Y_i$ , provided it is not "too" convex.

<sup>8</sup> See Hirshleifer, Section V.

<sup>9</sup> In fact, Theorem 1 shows that this conclusion holds regardless of the  $\Pi_i$ . Hence, even when tastes and prices are certain, the state-preference approach suggests that one cannot expect individuals to be risk averse in the usual sense at all times, since incomes are almost always uncertain and thus differ in the various states of nature.

cepted by this individual, even though  $V$  is strictly quasiconcave.

The line  $T$  is simply the tangent to  $V_0$  at the no-gamble point; it is the locus of  $(w_1, w_2)$  points satisfying

$$(5) \quad \Pi_1 w_1 U_y^1(Y_1, P) + \Pi_2 w_2 U_y^2(Y_2, P) = 0,$$

while the equation of the line  $F$  is just

$$(6) \quad \Pi_1 w_1 + \Pi_2 w_2 = 0$$

The assumption that the  $U^i$  are twice-differentiable functions of the  $Y_i$  implies that  $V_0$  cannot have a kink at the origin;  $w_1$  is a differentiable function of  $w_2$  along  $V_0$ . Hence, unless the lines  $F$  and  $T$  coincide, so that  $F$  is tangent to  $V_0$  at the no-gamble point, there exist unfair gambles which the individual would accept. Comparing equations (5) and (6), a necessary and sufficient condition that these lines coincide is that the marginal utility of income in the two states be the same. By considering in this fashion all possible gambles involving two states, three states, and so on, we obtain the main result of this section.<sup>10</sup>

**THEOREM 1:** *An individual is risk averse at the point  $([Y_i], P)$  if, and only if,  $U_y^i(Y_i, P) = U_y^j(Y_j, P)$  for all  $i$  and  $j$ .*

That is, an individual is risk averse at some point if, and only if, all marginal utilities of income are equal at that point. For any given  $Y_1$  and fixed price system  $P$ , there exists exactly one set of  $Y_2, \dots, Y_N$  for which any particular individual would be risk averse: that set which equates the

other marginal utilities to  $U_y^1(Y_1, P)$ . Thus the statement that an individual is risk averse at some point serves to relate the conditional utility functions to each other.

## II. No Contingent Claims Markets

In this section, it is assumed that there exist no markets for contingent claims of the sort discussed in much of the recent literature on behavior under uncertainty.<sup>11</sup> That is, we assume that the individual under study cannot contract in a market to have income paid him in any particular state of nature. Only an incomplete set of such markets exists in real economies, as discussed in Section IV. This section shows that option value, as defined by equation (1), may be positive, negative, or zero.

The formal results may be more easily understood if prefaced by an example of their implications. Consider the decision to save land with valuable industrial uses as a recreational facility. Many people would argue that the benefits of the recreational alternative always exceed the expected value of the consumer's surpluses it would generate in the various states of nature. They would point out that if the land is not saved, there is a risk that demand for recreation would be much higher than expected in the future, and so the foregone consumer's surplus would be large. They would contend that a risk premium, option value, should be added to expected consumer's surplus in order to compute the true benefits.

But this argument neglects the risk inherent in the other alternative. If the land is saved, there is a risk that demand for recreation will be much lower than expected and society will have given up the commercial use in exchange for very little. Both land use alternatives are risky, and which is socially riskier depends on the

<sup>10</sup> Alternatively, we could prove the theorem directly as follows. Consider the set of  $w_i$  such that  $V([Y_i + w_i], P) \geq V([Y_i], P)$  for some  $[Y_i]$ . By quasiconcavity, this set is convex and hence it has a supporting hyperplane at the point  $w_1 = w_2 = \dots = w_N = 0$ . By the assumption of twice-differentiability, the supporting hyperplane is unique, so that any other hyperplane passing through this point must also contain points in the interior of the set of  $w_i$  defined above. Examination of the generalizations of (5) and (6) in the text completes the proof.

<sup>11</sup> See Arrow (1964), Debreu, ch. 7, and William Brainard and F. Trenery Dolbear.

detailed structure of individuals' utility functions and contingent incomes. If the commercial use is riskier, option value is positive, but if the recreational use is riskier, option value is negative. There is no way a priori to specify the sign, let alone the magnitude, of option value, the net risk premium.

It is now necessary to define consumer's surplus, option price, and option value precisely. We deal first with *equivalent* measures. The equivalent consumer's surplus (the Hicksian "equivalent variation"),  $SE_i$ , generated in the  $i$ th state by a switch from  $P$  to  $P^*$ , assuming the individual's income in state  $i$  is  $Y_i$ , is defined by

$$(7) \quad U^i(Y_i, P^*) = U^i(Y_i + SE_i, P) \\ (i = 1, \dots, N)$$

That is,  $SE_i$  is the income gain in the  $i$ th state which is equivalent in that state to a change from  $P$  to the preferred  $P^*$ .

The equivalent option price,  $OPE$ , is the amount of income that would have to be given the same individual in every state of nature if  $P$  prevails in order to make him indifferent between  $P^*$  and  $P$ . Formally, the definition is

$$(8) \quad \sum_{i=1}^N \Pi_i U^i(Y_i, P^*) = \\ \sum_{i=1}^N \Pi_i U^i(Y_i + OPE, P)$$

Equivalent option value,  $OVE$ , can now be defined, following (1), by

$$(9) \quad OVE = OPE - \sum_{i=1}^N \Pi_i SE_i$$

We now prove

**THEOREM 2:** *If an individual with conditional incomes  $[Y_i]$  is risk averse at  $([Y_i + OPE], P)$ , and if the  $SE_i$  are not the same for all  $i$ ,  $OVE$  is nonpositive.*

PROOF:

Substitute for the left-hand side of (8) from (7) to obtain

$$(10) \quad \sum_{i=1}^N \Pi_i U^i(Y_i + SE_i, P) = \\ \sum_{i=1}^N \Pi_i U^i(Y_i + OPE, P)$$

Clearly if the  $SE_i$  are all equal they must equal  $OPE$ , and  $OVE$  is zero. Assume now that they are not all equal. Since the conditional utility functions are concave in income,

$$(11) \quad U^i(Y_i + SE_i, P) \leq U^i(Y_i + OPE, P) \\ + (SE_i - OPE) U_y^i(Y_i + OPE, P) \\ (i = 1, \dots, N)$$

Substituting (11) into (10) and subtracting equal terms from both sides of the resultant inequality, we obtain

$$(12) \quad \sum_{i=1}^N \Pi_i U_y^i(Y_i + OPE, P) (SE_i - OPE) \geq 0$$

The marginal utilities are equal by the assumption of risk aversion at  $([Y_i + OPE], P)$  and Theorem 1, and, from (9),  $OVE$  is nonpositive.

There is, of course, no particular reason to assume that the individual is risk averse at the point chosen in this theorem. Suppose instead that he is risk averse at  $([Y_i + SE_i], P)$ . (If the conditional utility functions are strictly concave, he cannot be risk averse both  $([Y_i + OPE], P)$  and  $([Y_i + SE_i], P)$ .) The same sort of reasoning employed to prove Theorem 2 establishes:

**THEOREM 3:** *If an individual with conditional incomes  $[Y_i]$  is risk averse at  $([Y_i + SE_i], P)$ , and if the  $SE_i$  are not the same for all  $i$ ,  $OVE$  is nonnegative.*

We now turn to *compensating magni-*

tudes. The compensating consumer's surplus (the Hicksian "compensating variation"),  $SC_i$ , generated in the  $i$ th state of nature by a switch from  $P$  to  $P^*$ , assuming conditional incomes  $[Y_i]$ , is defined by

$$(13) \quad U^i(Y_i - SC_i, P^*) = U^i(Y_i, P) \quad (i = 1, \dots, N)$$

That is, an income loss of  $SC_i$  in the  $i$ th state of nature compensates for a shift from  $P$  to  $P^*$  in the sense of leaving the individual exactly as well off in that state as he was before the shift.

The compensating option price,  $OPC$ , is defined by

$$(14) \quad \sum_{i=1}^N \Pi_i U^i(Y_i - OPC, P^*) = \sum_{i=1}^N \Pi_i U^i(Y_i, P),$$

and the compensating option value,  $OVC$ , is given, following (1), by

$$(15) \quad OVC = OPC - \sum_{i=1}^N \Pi_i SC_i$$

Again using the same sort of reasoning employed in the proof of Theorem 2, we can establish the following results:

**THEOREM 4:** *If an individual with conditional incomes  $[Y_i]$  is risk averse at  $([Y_i - OPC], P^*)$ , and if the  $SC_i$  are not the same for all  $i$ ,  $OVC$  is nonnegative.*

**THEOREM 5:** *If an individual with contingent incomes  $[Y_i]$  is risk averse at  $([Y_i - SC_i], P^*)$ , and if the  $SC_i$  are not the same for all  $i$ ,  $OVC$  is nonpositive.*

If the  $U^i$  are strictly concave functions of the  $Y_i$ , Theorems 2–5 become somewhat stronger; "positive" replaces "nonnegative" in Theorems 3 and 4, and "negative" replaces "nonpositive" in Theorems 2 and 5.

It should be clear that option value is not necessarily positive. If an individual is

not risk averse at any of the points considered in Theorems 2–5, the sign of option value cannot be determined in general. If an individual is always risk averse when his income is certain, if the  $U^i$  are strictly concave, and if all the  $Y_i$  are equal,  $OVC$  is negative and  $OVC$  is positive—compare Theorems 2 and 4. If the  $U^i$  are the same but the  $Y_i$  differ, both option values may be positive or negative.

Since it can be positive, negative, or zero, depending on the measure chosen and the structure of preferences, one might suspect that option value is a second-order magnitude that always vanishes for small price changes. This conjecture is false, however, as is now demonstrated. Assume that the  $U^i$  are differentiable functions of the elements of the price vector, and write that vector as  $\alpha P^* + (1 - \alpha)P$ .

For small changes in  $\alpha$  from zero to  $d\alpha$ , the two surplus measures defined by (7) and (13) reduce to  $dS_i$ , defined by

$$(16) \quad U_y^i(Y_i, P)dS_i + U_\alpha^i(Y_i, P)d\alpha = 0 \quad (i = 1, \dots, N)$$

Similarly, the two measures of option price defined by (8) and (14) reduce to  $dOP$ , which satisfies

$$(17) \quad \sum_{i=1}^N \Pi_i [U_y^i(Y_i, P)dOP + U_\alpha^i(Y_i, P)d\alpha] = 0$$

The corresponding option value,  $dOV$ , is given by

$$(18) \quad dOV = dOP - \sum_{i=1}^N \Pi_i dS_i$$

Substituting (16) into (17) and solving yields

$$(19) \quad dOP = \frac{\sum_{i=1}^N \Pi_i U_y^i(Y_i, P) dS_i}{\sum_{i=1}^N \Pi_i U_y^i(Y_i, P)}$$

Thus in the case of small price changes, option price can easily be expressed as a weighted average of the conditional surpluses. Clearly, though, the weights equal the probabilities in general if, and only if, the marginal utilities of income are equal in all states. Recalling Theorem 1, we thus have:

**THEOREM 6:** *If an individual with contingent incomes  $[Y_i]$  is risk averse at  $([Y_i], P)$ , the option value associated with small price changes,  $dOV$ , is zero. If he is not risk averse at this point,  $dOV$  may be positive or negative.*

The general principle underlying these results may be illustrated as follows. Consider two sets of conditional incomes,  $[Y_i]$  and  $[Y'_i]$ , both with the same expected value. If an individual is risk averse at  $([Y_i], P)$ , a shift to the situation  $([Y'_i], P)$  represents a fair gamble and would not be accepted voluntarily. The first income set is preferred, and a risk premium must be paid in order to induce a switch to the second. Similarly, risk aversion at the second point implies that it is preferred to the first. When the  $U^i$  differ, it must be specified just how they differ before it can be determined which of two sets of conditional incomes with the same expected value is preferred.

### III. Complete Contingent Claims Markets

In this section we assume that perfectly competitive markets exist for contingent income claims in all states of nature. In such markets, any individual can contract for income to be paid him if, and only if, any particular state of nature occurs. Let  $q_i$  ( $i=1, \dots, N$ ) be the market price for one dollar to be paid if state  $i$  occurs. In the timeless model we are considering, the sum of the  $q_i$  across all states must be unity, since by purchasing one dollar in each state an individual buys one dollar of

certain income.<sup>12</sup> There is, however, no general presumption that the  $q_i$  equal the  $\Pi_i$ , the probabilities of the various states.<sup>13</sup> We assume that the individual has a set of pretrade conditional incomes,  $[Y_i^0]$ , the market value of which we call his initial endowment. As before, he is assumed to spend this endowment so as to maximize his utility, which is given by equation (3).

Suppose that the price system  $P$  prevails in all states of nature. Given his endowment,  $E'$ , the individual chooses a set of posttrade conditional incomes  $[Y'_i]$  so as to maximize (3) subject to the budget constraint

$$(20) \quad \sum_{i=1}^N q'_i Y_i^0 \equiv E' = \sum_{i=1}^N q'_i Y'_i,$$

where the  $q'_i$  are the contingent claims prices which result if price system  $P$  prevails. Denote the (maximum) level of expected utility thus attained by  $V'$ .

Now consider the situation where the preferred price system  $P^*$  prevails in all states and the initial endowment is  $E^*$ . Call the (maximum) level of expected utility attained  $V^*$ , and let  $Y_i^*$  be the income the individual contracts to receive in the  $i$ th state ( $i=1, \dots, N$ ). The  $Y_i^*$  must satisfy the budget constraint

$$(21) \quad \sum_{i=1}^N q_i^* Y_i^0 \equiv E^* = \sum_{i=1}^N q_i^* Y_i^*$$

where the  $q_i^*$  are the  $q_i$  corresponding to price system  $P^*$ .

The equivalent option price,  $OPE$ , is most sensibly defined in this context as the amount of additional initial endowment which must be given to the individual if

<sup>12</sup> See Arrow (1964) for a formal discussion. If time is introduced into the model and if the  $q_i$  relate to states of nature one period in the future, their sum is the reciprocal of one plus the one-period riskless rate of interest: see Hirshleifer, and Arrow and Robert Lind.

<sup>13</sup> On relations between the  $q_i$  and the  $\Pi_i$  in different contexts, see Arrow and Lind, and Brainard and Dolbear.

price system  $P$  prevails in order to make  $V'$  equal to  $V^*$ . Similarly, the compensating option price,  $OPC$ , is logically the amount by which  $E^*$  must be reduced if the price system  $P^*$  prevails in order to lower  $V^*$  to equality with  $V'$ .

It is less obvious how to define the two measures of consumer's surplus here. Definitions (7) and (13) cannot be used directly when markets for contingent claims exist, since income can be shifted from state to state, and this leaves the  $V_i$  in those equations ambiguous. In fact, there would seem to be no sensible way to associate a unique surplus with any particular state of nature, and thus no way to associate a unique set of surpluses with any price change.

Still, it is easy to prove that the market value of any such set must equal the corresponding option price. This is little more than the statement that if markets are perfect and complete, the market value of the whole must equal the sum of the market values of its parts, as we now show.

Consider first the equivalent measures. Suppose the consumer is indifferent between price system  $P^*$  and endowment  $E^*$  on the one hand and price system  $P$ , endowment  $E'$ , and any set of conditional payments  $[SE_i]$  on the other hand. Then  $[SE_i]$  can be interpreted as one set (of an infinite number of sets) of equivalent consumer's surpluses. In the second situation, the consumer's endowment is

$$(22) \quad \sum_{i=1}^N q_i' SE_i + E'$$

By assumption, he is also indifferent between  $P^*$  with endowment  $E^*$  and  $P$  with endowment  $E' + OPE$ . Hence, it must be the case that (22) equals  $E' + OPE$ , and

$$(23) \quad \sum_{i=1}^N q_i' SE_i = OPE$$

Similarly, if the consumer is indifferent between price  $P$  and endowment  $E'$  on the

one hand and price system  $P^*$ , endowment  $E^*$ , and a set of conditional payments  $[-SC_i]$  on the other, the market value of the set of the  $CS_i$ , computed using the  $q_i^*$ , must equal  $OPC$ . Thus when a complete set of markets for contingent claims exists, option price is nothing more or less than the market value of a corresponding set of conditional consumer's surpluses, making the two measures of the value of a price change essentially the same.

The market value of a set of conditional surpluses in general equals its expected value if, and only if, the prices of conditional payments, the  $q_i$ , are equal to the corresponding probabilities, the  $\Pi_i$ . When this occurs, option value as defined by equation (1) is zero. It is thus of some interest to ask when the  $q_i$  will equal the  $\Pi_i$ ; when an individual will be allowed to trade at prices equal to the probabilities.

Sufficient conditions for this can be obtained by modifying a discussion of Arrow and Lind.<sup>14</sup> Consider a group, small relative to society, facing uncertainty about their pretrade incomes (the  $Y_i^0$  above) and tastes. Assume that all probabilities are objective (i.e., universally agreed upon), that prices do not depend on the state of nature, and that tastes and pretrade incomes of the group members are determined independently (in the statistical sense) of those of the rest of the society. We now indicate that these conditions guarantee the ability of the group members to trade in claims conditional on their tastes and pretrade incomes at prices equal to the relevant probabilities.

It is convenient to partition the possible states of nature in two ways. First, there exist mutually exclusive and collectively exhaustive sets of states of nature  $Q^g$  ( $g=1, \dots, G$ ) such that for all states in any given  $Q^g$  the tastes and pretrade incomes of each individual in the group are

<sup>14</sup> Arrow and Lind, Section I.

the same. (Loosely, this partition captures all distinctions directly relevant to the group members.) Similarly, there exist mutually exclusive and collectively exhaustive sets of states  $Q_s$  ( $s=1, \dots, S$ ) such that for all states in any given  $Q_s$  the tastes and pretrade incomes of each individual in the rest of society are the same. Let  $\Pi_g$  be the probability that a state in  $Q^g$  will occur, and let  $\Pi_s$  be the probability that a state in  $Q_s$  will occur. The assumption of independence then implies that  $\Pi_{gs}$ , the probability that a state will lie in both  $Q^g$  and  $Q_s$  is equal, for all  $g$  and  $s$ , to  $\Pi_g \Pi_s$ , and the assumption of objectivity implies that these probabilities are agreed upon by all members of society. States in both  $Q^g$  and  $Q_s$  will be said to lie in the set  $Q^g_s$ .

Consider the situation before the small group trades in contingent claims with the rest of society. Every individual in the rest of society will contract to receive the same income in every state of nature in each  $Q_s$ , for these states are effectively indistinguishable before trade with the small group. Now suppose that the small group is so tiny relative to society that trade with it has a negligible impact on the prices of contingent claims and hence on the conditional incomes contracted for by members of the rest of society.<sup>15</sup> Then the equilibrium posttrade income of a typical individual in the rest of society in any state depends only on which of the  $Q_s$  that state is in.

The equilibrium conditions for a typical member of the rest of society are<sup>16</sup>

$$(24) \quad \Pi_g \Pi_s U_y^s(Y_s, P) = \lambda q_{gs} \\ (g = 1, \dots, G; s = 1, \dots, S),$$

where  $Y_s$  is the income he contracts to receive in all states in  $Q_s$ ,  $q_{gs}$  is the market price of a dollar of income to be paid if the state of nature lies in  $Q^g_s$ , and  $\lambda$  is a Lagrange multiplier. Comparing the equations in (24) that correspond to states in the same  $Q_s$  but different  $Q^g$ , it is easy to show that

$$(25) \quad q_{gs} = \Pi_g q_s \\ (g = 1, \dots, G; s = 1, \dots, S),$$

where  $q_s$  is the price of a contract which pays one dollar in all states in  $Q_s$ . Since the partition of states into the  $Q_s$  is collectively exhaustive,  $q_g$ , the price of a contract which pays one dollar in all states in  $Q^g$ , is given by

$$(26) \quad q_g = \sum_{s=1}^S q_{gs} = \Pi_g \sum_{s=1}^S q_s = \Pi_g$$

From (26), members of the small group are able to trade claims contingent on their pretrade incomes and tastes at prices which equal the relevant probabilities. Hence the option value associated with price changes which only affect this group, using either measure of option value, is zero. Given (26), the equilibrium conditions for a typical group member are analogous to (24):

$$(27) \quad \Pi_s U_y^g(Y_s, P) = \mu q_s \\ (g = 1, \dots, G; s = 1, \dots, S),$$

where  $\mu$  is a Lagrange multiplier. Note that we cannot assume in general that this individual contracts to receive the same income in all states in each  $Q^g$ . If  $\Pi_s = q_s$ , however,  $Y$  can depend only on  $g$ . Further, if this condition holds, this individual has equal marginal utilities of income in all states of nature and is thus (from Theorem 1) risk averse at equilibrium.

<sup>15</sup> The assumption of negligibility is also important in the Arrow-Lind argument, though their interpretation of their formal result (p. 369, below equation 13) gives the impression that it is not. The investment they consider must have a negligible impact on society's income in all states of nature, since otherwise individuals' allocations (the  $x_{iag}$  on p. 369) will not be independent of its success or failure, and their equation (13) will not hold.

<sup>16</sup> See Arrow and Lind, Section I.

If every member of society is a member of a small group such as the one we have singled out for consideration, every member of society is able to trade claims contingent on his pretrade income and tastes at prices equal to the relevant probabilities. Then the option value associated with any price change is zero, and all individuals are risk averse at equilibrium.

The assumption that society consists of many small groups of consumers whose tastes are determined independently does not seem overly restrictive. The assumption of independent income determination, however, is quite strong. In no society is total income certain, nor are the incomes of members (or relevant groups of members) uncorrelated, since shares of national income typically change less rapidly than national income itself varies over the business cycle.<sup>17</sup>

#### IV. The Real World

In Section II we found that if no markets for claims contingent on incomes and tastes exist, option price may exceed or fall short of expected consumer's surplus. In Section III, on the other hand, we showed that if a complete set of perfect contingent markets exists, option price is just the market value of the relevant contingent consumer's surpluses. We then gave conditions under which option price equals the expected value of these surpluses. In this final section we briefly attempt to relate these abstract results to the real world.

A number of authors have noted that a complete set of markets for claims contingent on incomes do not exist in real economies, and they have discussed the reasons for this.<sup>18</sup> There are, in addition, several special reasons why markets for claims contingent on individuals' tastes are never

(to my knowledge) observed.<sup>19</sup> First, it is difficult to make an objective judgement about the probabilities associated with the evolution of one's own tastes, let alone those of others. Second, the problem of moral hazard arises in a severe form in such markets.<sup>20</sup> The evolution of my preferences is not likely to be independent of gambles I make regarding my future tastes. That is, the relevant probabilities will change when an individual purchases or sells claims contingent on his future tastes.

Finally, one individual cannot generally ascertain another's tastes directly; he can at best infer them from choices made in the marketplace. If an individual has signed contracts for payments contingent on his tastes, and if his tastes can only be judged by his market behavior (as opposed to reading his thoughts), this will affect that behavior and make it difficult ever to know what contingent claims should be paid.

Let us now consider the problem faced by a government agency in a real economy that is deciding whether or not to undertake an investment which would noticeably lower a market price. Ideally, it would like to know the sum of individuals' option prices corresponding to that change, as these are "certainty equivalent" magnitudes. If the price change is to occur in the future, this sum should simply be discounted at the riskless rate of interest.

In the absence of complete and perfect markets for risk bearing, option prices are

<sup>19</sup> The market for health care insurance might be interpreted as being of this type, but even this is not a pure case, as such insurance usually entitles one to care only if a doctor feels care is needed. Thus I cannot have my perfectly healthy appendix removed at the insurance company's expense even though I would be glad to be rid of it, nor can I vacation for free in a hospital. No other examples of markets for claims contingent on individuals' tastes come to mind, though some may exist.

<sup>20</sup> See Mark Pauly and Arrow (1968) for discussions of moral hazard.

<sup>17</sup> See Brainard and Dolbear, Section II, on this point.

<sup>18</sup> See Brainard and Dolbear and the references there cited.

unfortunately unobservable. About the best one can expect is information about the consumer's surpluses that would be generated in the various states of nature, along with some estimates of the relevant probabilities. Should the expected values of these surpluses be used in place of option price as a benefit measure? Put another way, if the price change is to occur in the future, should the sum of these expected values be discounted at a rate above, below, or equal to the riskless rate of interest?

Arrow and Lind argue that when the gains from a project are uniformly and thinly distributed over the entire population, expected values (and hence riskless discount rates) are generally appropriate.<sup>21</sup> This argument requires the assumption of risk aversion, however, as Theorem 6 makes clear. The analysis of Sections I-III indicates that this is not a particularly compelling assumption in the state-preference framework. Further, as these authors also point out, benefits from government investments typically accrue mainly to a fraction of society, and risk-spreading arguments have little force in such cases.

The usual argument is that in these cases, benefits should be discounted at a rate above the riskless rate, just as the private market discounts risky income streams. But even this is not obvious. When social income is the main source of uncertainty, some idea of the appropriate rate might be gotten by valuing more highly (discounting less heavily) benefits which occur in states of nature where social income is low, since even incomplete contingent claims markets would value a dollar in recession more than a dollar in prosperity.<sup>22</sup> Thus if an investment yields benefits which are negatively correlated

with national income, those benefits might well be discounted at a rate below the riskless rate, while most investments, with returns positively correlated with the cycle, should have their benefits discounted at a rate above the riskless rate.

When uncertainty about tastes is important, though, no such rules of thumb are apparent. There are no markets for claims contingent on tastes from which one might get information. Individuals' option prices may exceed or fall short of the expected value of the contingent surpluses they would derive from a price change, and it is not obvious how one might judge in a real situation which was more likely. This suggests that when tastes are the main source of uncertainty, the expected value of consumers' surpluses ought to be employed as the best available approximation to the sum of their option prices, and this approximate total should be discounted at the riskless rate of interest. Benefits will be sometimes underestimated and sometimes overestimated with this procedure, but there would appear to be no practical way to obtain superior estimates.

#### REFERENCES

- K. J. Arrow, "The Role of Securities in the Optimal Allocation of Risk-Bearing," *Rev. Econ. Stud.*, Apr. 1964, 31, 91-96.
- , "The Economics of Moral Hazard: Further Comment," *Amer. Econ. Rev.*, June 1968, 58, 537-39.
- and A. C. Enthoven, "Quasi-Concave Programming," *Econometrica*, Oct. 1961, 29, 779-800.
- and R. C. Lind, "Uncertainty and the Evaluation of Public Investment Decisions," *Amer. Econ. Rev.*, June 1970, 60, 364-78.
- W. Brainard and F. T. Dolbear, "Social Risk and Financial Markets," *Amer. Econ. Rev. Proc.*, May 1971, 61, 360-70.
- D. R. Byerlee, "Option Demand and Consumer Surplus: Comment," *Quart. J. Econ.*, Aug. 1971, 85, 523-27.

<sup>21</sup> Arrow and Lind, Sections II and III.

<sup>22</sup> See Brainard and Dolbear, Section II, especially fn. 9 of their (unpublished) complete paper, on this point.

- D. Cass, "Discussion," *Amer. Econ. Rev. Proc.*, May 1969, 59, 562-63.
- C. J. Cicchetti and A. M. Freeman II, "Option Demand and Consumer Surplus: Further Comment," *Quart. J. Econ.*, Aug. 1971, 85, 528-39.
- G. Debreu, *Theory of Value*, New York 1959.
- J. Hirshleifer, "Investment Decision Under Uncertainty: Choice-Theoretic Approaches," *Quart. J. Econ.*, Nov. 1965, 79, 509-36.
- J. V. Krutilla, C. J. Cicchetti, A. M. Freeman III, and C. S. Russel, "Observations on the Economics of Irreplaceable Assets," in A. V. Kneese and B. T. Bower, eds., *Environmental Quality Analysis: Theory and Method in the Social Sciences*, Baltimore 1972.
- C. M. Lindsay, "Option Demand and Consumer's Surplus," *Quart. J. Econ.*, May 1969, 83, 344-46.
- M. F. Long, "Collective-Consumption Services of Individual-Consumption Goods: Comment," *Quart. J. Econ.*, May 1967, 81, 351-52.
- J. von Neumann and O. Morgenstern, *The Theory of Games and Economic Behavior*, 2d ed., Princeton 1947.
- M. V. Pauly, "The Economics of Moral Hazard: Comment," *Amer. Econ. Rev.*, June 1968, 58, 531-37.
- B. A. Weisbrod, "Collective-Consumption Services of Individual-Consumption Goods," *Quart. J. Econ.*, Aug. 1964, 78, 471-77.
- R. Zeckhauser, "Resource Allocation with Probabilistic Individual Preferences," *Amer. Econ. Rev. Proc.*, May 1969, 59, 546-52.

# Sectoral Investment Determination in a Developing Economy

By JERE R. BEHRMAN\*

The dominant point of focus in most macroanalyses of economic development has been the accumulation of real physical capital. This dominance is most explicit in many formal models in which output is postulated to depend only on the stock of real physical capital.<sup>1</sup> In recent years this concentration on the role of real physical capital in development has come under question because of empirical evidence that other factors have been important, that the elasticities of substitution between real physical capital and other factors are significantly nonzero, and that production capacities (whether created by real physical capital alone or in combination with other factors) have been significantly underutilized.<sup>2</sup> Nevertheless, the accumulation of real physical capital is

still widely thought to be a very important factor in economic development. The question of what determines real physical capital investment in the developing economies, therefore, is a very important one. The present study reports on attempts to estimate sectoral real physical capital investment functions from time-series data for the developing economy of postwar Chile. In addition to insights gained about real investment determinants in a developing country, this study may contribute to the more general controversy over investment determination carried on in recent years in this *Review*.<sup>3</sup> The models utilized are presented in Section I. The results obtained are presented and discussed in Section II. Concluding remarks are made in Section III.

## I. Models of Real Physical Capital Investment Behavior

Both "putty-putty" and "putty-clay" investment models are explored in this study.<sup>4</sup> The putty-putty models are further subdivided into two alternatives depending upon the nature of real replace-

\* Professor of economics, University of Pennsylvania. The research on this publication was supported by a Ford Foundation grant to the Office of Chilean National Economic Planning (ODEPLAN) and to the Center for International Studies of the Massachusetts Institute of Technology and by a Ford Foundation Faculty Fellowship in Economics. The research was initiated while I was a resident economic investigator at ODEPLAN in 1963-69. I wish to thank the supporting organizations for their help, and at the same time emphasize that the conclusions, opinions and other statements in this publication in no way necessarily reflect the viewpoints of these organizations. I also wish to thank, but not implicate, Edmar Bacha, Juan de la Barra, Charles Bischoff, Peter Clark, Eduardo García, Lawrence Klein, Ricardo Lira, Jozé Mencinger, Christian Ossa, and Lance Taylor.

<sup>1</sup> For some recent examples, see Irma Adelman and Frederick Sparrow, Hollis Chenery and Alan Strout and United Nations Conference on Trade and Development (and J. Behrman (1972a) for a review of the last study).

<sup>2</sup> For such empirical evidence in the case of the developing country for which empirical results are presented in this study (i.e., Chile), see Behrman (1973a,b), (1972b,c) and Arnold Harberger and Marcelo Selowsky.

<sup>3</sup> See Charles Bischoff (1969), (1971), Robert Coen (1968), (1969), (1971), Robert Eisner (1960), (1967), (1969a,b), Eisner and M. Ishaq Nadiri, Eisner and Robert Strotz, Franklin Fisher (1971), Robert Hall and Dale Jorgenson (1967), (1969), (1971), Jorgenson (1963), (1965), Jorgenson and Calvin Siebert, Jorgenson and James Stephenson (1967a,b), and Lawrence Klein and Paul Taubman.

<sup>4</sup> Models dependent on liquidity also were explored, but—perhaps due to substantial data problems—the results are of little interest and are not presented here. For discussions of the fairly strong assumptions in respect to aggregation and in respect to perfect competition which are required for the putty-putty and putty-clay models, see Fisher (1971) and Klein and Taubman.

ment investment. If the same opportunities are available for real replacement investment as for real net investment, there is no effective distinction between the two and the model is called "complete putty-putty." If the opportunities available for real replacement investment are more limited than are those for real net investment due to fixed characteristics of surviving capital (or of other factors) in the same process, there is a distinction between real replacement and real net investment, and the model is called "partial putty-putty."

Under general putty-putty assumptions, real gross physical capital investment is the sum of real net investment and real replacement investment:<sup>5</sup>

$$(1) \quad I^G = I^N + I^R$$

where

$I^G$  = real gross physical capital investment

$I^N$  = real net physical capital investment

$I^R$  = real replacement physical capital investment

Real net physical capital investment is hypothesized to depend upon changes in the desired real physical capital stock with a distributed lag adjustment over  $n+1$  periods because of lags in the decision-making and implementation processes:<sup>6</sup>

$$(2) \quad I^N = \sum_{i=0}^n a_i \Delta K_{t-i}^D$$

where

$K^D$  = desired real physical capital stock.

One would expect the sum of the weights

<sup>5</sup> To simplify the notation, time subscripts and disturbance terms are omitted unless they are needed to prevent confusion.

<sup>6</sup> In the case of the particular country under examination in which the mean annual rate of change of the gross domestic product deflator in the sample period was 30 percent, these lags probably reflect in part the expectation formation aspect of the decision-making process for the relative prices included in relations (3), (3a), (5), and (6) below.

in the lag adjustment to be one if any change in desired real physical capital stock is eventually to result in net investment of exactly the same magnitude:

$$(2a) \quad \sum_{i=0}^n a_i = 1$$

The desired real physical capital stock is assumed to depend primarily on neo-classical investment behavior.<sup>7</sup> In contrast to the assumption of Jorgenson and his various collaborators, however, the underlying technology is not assumed to be only Cobb-Douglas. Instead, the more general assumption is made of a CES production function with constant returns to scale and with Hicks neutral technological change:<sup>8</sup>

$$(3) \quad K^D = b \left( \frac{P_Q}{P_K} \right)^{\sigma} Q e^{r(\sigma-1)t}$$

where

$P_Q$  = price of output

$P_K$  = price of capital services<sup>9</sup>

$Q$  = real output (value-added)

$\sigma$  = elasticity of substitution between capital and labor

$r$  = rate of Hicks neutral exponential technological change

$b$  = constant<sup>10</sup>

<sup>7</sup> See references in fn. 3 to articles by Jorgenson and various collaborators.

<sup>8</sup> Bischoff (1969), (1971) and Coen (1968), (1969), (1971) also start with the more general CES production function instead of the special Cobb-Douglas case. The seminal article on the CES production function is, of course, by Kenneth Arrow et al. For excellent and more recent relevant discussions, see Murray Brown (1966), (1967) and Fisher (1969).

<sup>9</sup> The price of capital services depends on the price of investment goods, the cost of capital, and the tax structure. See Jorgenson and Siebert, p. 695. In the present study, capital gains are not included explicitly in the price of capital services, because the cost of capital series utilized better approximates the real than the nominal cost of capital and thus represents the general effect of inflation (see fn. 6 above). The impact of sector-specific price deviations from the general inflation is included in the last term in relation (3a).

<sup>10</sup> The constant depends on the elasticity of substitution between capital and labor, the scale parameter ( $\gamma$ )

If the elasticity of substitution between capital and labor is one, this formulation reduces to that used by Jorgenson. If the elasticity of substitution between capital and labor is zero, this formulation reduces to the accelerator model used by Eisner and others.<sup>11</sup> In the present study this formulation is used with three values of the elasticity of substitution for each data cell: zero, an estimate from an earlier study (between zero and one in value),<sup>12</sup> and one. The results, thus, may provide some insight into the sensitivity of investment behavior estimates to the range of assumptions about the value of the elasticity of substitution between capital and labor which underlies much of the Jorgenson-Eisner debate. Possible modifications of the desired capital stock due either to underutilization of capacity<sup>13</sup> or to variance in the product price relative to the overall price level<sup>14</sup> also are explored by the addition of linear terms to represent such considerations:

$$(3a) \quad K^D = b \left( \frac{P_Q}{P_K} \right)^\sigma Q e^{r(\sigma-1)t} + c \frac{Q}{Q^e} \\ + dSD \left( \frac{P_Q}{P_{GDP}} \right)$$

which denotes the initial efficiency of the technology, and the capital intensity parameter ( $\delta$ ):

$$b = \gamma^{\sigma-1} \delta^\sigma$$

<sup>11</sup> See references in fn. 3 to articles by Eisner and various collaborators.

<sup>12</sup> See Behrman (1972b). The estimates for the exponential rate of Hicks neutral technological change are also taken from this same source.

<sup>13</sup> Capacity is defined by the trend-through-the-peaks procedure suggested by Klein and Robert Summers. For critical evaluations of this procedure also see Behrman (1973a), Almarin Phillips (1963), (1969), and Summers. See Behrman (1973a) for an examination of the extent and cause of capacity underutilization for data cells used in this study.

<sup>14</sup> The determination of overall Chilean price changes is explored in Behrman (1970c), Peter Gregory, Harberger, and Joseph Ramos. The determination of price changes for the data cells of this study is explored in the first of these studies.

where

$Q^e$  = the capacity of real output (value-added)

$SD(X)$  = standard deviation of  $X$  over three periods ( $t, t-1, t-2$ )

$P_{GDP}$  = deflator for gross domestic product

If the degree of capacity utilization is relevant in the real desired capital decision, the coefficient of the first additional variable should be positive and significantly nonzero. If expectations of the standard deviation of product prices relative to the overall price level are relevant and are represented sufficiently well by the above formulation, and if risk aversion is predominant, the coefficient of the second additional variable should be negative and significantly nonzero. Finally, when relation (3a) is substituted back into relation (2), the possibility of differential patterns of lagged adjustment to the different components of the desired real physical capital stock is allowed in the estimation below. Considerable flexibility is permitted in the lag specification by allowing the lag pattern to be represented by polynomials of up to the fourth degree over a four-year period (i.e.,  $n=3$ ) in the method suggested originally by Hall and Robert Sutch.<sup>15</sup>

In the general putty-putty model real replacement physical capital investment is represented according to two alternative considerations. First, if capital (and other factors) is really malleable (i.e.,

<sup>15</sup> In the present study the "tail" included by Hall and Sutch is constrained to zero. With such a constraint the Hall-Sutch technique is formally identical to the Almon technique, except that only the right-hand tail is constrained to zero and that the various terms included in the procedure represent coefficients of the linear, quadratic, cubic, and/or quartic terms in the lag polynomial. Given the assumption of adjustment in four years (i.e.,  $n$  in relation (2) is set equal to three), of course, if all four terms are included in the lag polynomial, the procedure utilized is equivalent to including the contemporary variable and the first three lagged values of this variable with no restrictions on the form of the lag polynomial.

complete putty-putty), then it would seem to make no difference whether gross investment were for replacement or for net capital stock addition purposes. In either case the same range of alternatives would be available, and the formulation discussed above for real net physical capital investment would seem to be equally appropriate for real replacement physical capital investment. In such a case one could not distinguish between the determinants of net and of replacement real physical capital investment. Second, if replacement options are limited due to fixed characteristics of surviving capital or of other factors in the same process (i.e., partial putty-putty or putty-putty in net investment but not in replacement investment), replacement needs may be determined directly by the depreciation of the existing real physical capital stock.<sup>16</sup> Due to lags in the decision-making and implementation processes, once again real replacement physical capital investment in any one period may depend on the needs so generated for several ( $m$ ) periods (perhaps as modified by the state of technology for the period in which the real replacement physical capital investment actually occurs). Because of the very limited information available about the size of the capital stock at any point in time for the data cells of the present study,<sup>17</sup> the capacity of real output (corrected for the state of technology at the time the investment occurs by the same factor as is included in

relation (3) above) is used as a proxy for the real physical capital stock:

$$(4) \quad I^R = \left( \sum_{i=0}^m f_i Q_i^c \right) e^{r(\sigma-1)t}$$

For estimation purposes in this second alternative, relation (4) is substituted into relation (1) together with the appropriate representation of real net physical capital investment. The lag structure for real replacement physical capital investment also is represented by the Hall-Sutch polynomial method over a four-year period (i.e.,  $m=3$ ).

Under general putty-clay assumptions (in terms of the notation presented above), real gross physical capital investment depends on the price of output, the price of capital services, real output, the elasticity of substitution between capital and labor, the rate of Hicks neutral exponential technological change, and the rate of depreciation in the following manner:<sup>18</sup>

$$(5) \quad I^G = b e^{r(\sigma-1)t} \left[ \sum_{i=0}^{n_1} \left\{ w_{1i} \left( \frac{P_Q}{P_K} \right)_{-i} \right\}^\sigma \right] \cdot \left[ \sum_{j=0}^{n_2} w_{2j} \{ Q_{-j} - (1-d)Q_{-j-1} \} \right]$$

where  $w_{1i}$  and  $w_{2j}$  are weights for the two sums, and  $d$  is the yearly depreciation rate.

A first-order Taylor series expansion of relation (5) around the lagged values of  $P_Q/P_K$  and  $\{Q - (1-d)Q_{-1}\}$  gives relation (6) where bars are used to represent lagged values. Under a priori assumed values of the depreciation rates and the duration of the lag periods (as in the putty-putty models above, a four-year lag period is assumed so that  $n_1=n_2=3$ ), a two-dimensional search over values of the elasticity of substitution between capital and labor

<sup>16</sup> This is the procedure generally followed by Jorgenson and his collaborators in the studies referred to in fn. 3 above.

<sup>17</sup> In some experiments 1965 real physical capital stock values from the División de Programación Global, ODEPLAN, were used with assumptions about depreciation rates and the real physical capital investment series to generate a real physical capital stock series. The use of these series in production function estimates did not lead to a priori very satisfactory results.

<sup>18</sup> See the references to the studies by Bischoff in fn. 3 above. I gratefully acknowledge the contribution which Bischoff made to this discussion by suggesting the formulation presented here in place of an earlier incorrect presentation.

$$\begin{aligned}
 (6) \quad I^G = & be^{r(\sigma-1)t}(\overline{P_Q/P_K})^\sigma \{ \overline{Q} - (1-d)Q_{-1} \} \\
 & + \sum_{i=0}^{n_1} w_{1i} be^{r(\sigma-1)t} (\overline{P_Q/P_K})^{\sigma-1} \{ \overline{Q} - (1-d)Q_{-1} \} \{ (P_Q/P_K)_{-i} - (\overline{P_Q/P_K}) \} \\
 & + \sum_{j=0}^{n_2} w_{2j} be^{r(\sigma-1)t} (\overline{P_Q/P_K})^\sigma \{ [Q_{-j} - (1-d)Q_{-j-1}] - \{ \overline{Q} - (1-d)Q_{-1} \} \}
 \end{aligned}$$

( $\sigma=0.33, 0.67, 1.0, 1.33$ ) and over the rate of Hicks neutral exponential technological change ( $r=0.000, 0.005, 0.010, 0.015, 0.020, 0.025, 0.030, 0.040, 0.050$ ) is conducted.<sup>19</sup> Once again the Hall-Sutch polynomial method is used, and possible modifications of the investment decision due to underutilization of capacity or to variance in the product price relative to the overall price level are explored.

## II. Estimates of Real Physical Capital Sectoral Investment Behavior in Postwar Chile

Under the assumption that maximum likelihood procedures are appropriate (i.e., the additive disturbance terms are distributed normally with mean zero, with a diagonal variance-covariance matrix, with constant own variance, and independently of contemporaneous right-hand side variables), the models described in the previous section were estimated for the following six Chilean sectors for the 1945-65 period: agriculture, mining, manufacturing, transportation, utilities, and housing.<sup>20</sup> The ranges, means and standard

deviations of annual real gross physical capital investment for each of these sectors over the sample period are presented in Table 1. Also included in this table are the percentage of total real gross physical capital investment which occurred in each sector and the percentages of total product which originated in each sector. These data indicate that in order of the size of real gross physical capital investment the sectors are housing, manufacturing, transportation, utilities, agriculture, and mining. In order of the size of the standard deviation of real gross physical capital investment the sectors are transportation, housing, manufacturing, utilities, mining, and agriculture. If these standard deviations are normalized by the means, then this ordering changes to mining, utilities, agriculture, transportation, manufacturing, and housing. Thus housing, manufacturing, and transportation have accounted for a substantial proportion (75 percent over the sample) of the total real gross physical capital investment although at least in a relative sense the variances in the real gross physical capital investment in these three sectors have been small. For the foreign controlled (at least during the sample) mining sector, in contrast, the relative fluctuations in real gross physical capital investment have been rather large. Examination of the last two columns in Table 1 suggests that in

<sup>19</sup> Since this model reduces to the previously treated accelerator model when the elasticity of substitution between capital and labor is zero, that case is not included in this scanning. Nor is the estimate of the elasticity of substitution between capital and labor from Behrman (1972b) included since that estimate is based on putty-putty assumptions. The range for the elasticity of substitution between capital and labor is extended beyond the range indicated in the text, however, in all cells in which a value at the endpoint of the range in the text initially is most consistent with the variation in real gross physical capital investment.

<sup>20</sup> The estimation for transportation is for the 1950-65 period due to a lack of investment data prior to 1950.

The lack of comparable data before 1940 and after 1965 and the desire to exclude the war years due to the special considerations therein underlie the selection of the 1945-65 period for the other sectors. Data sources for all sectors are given in the Appendix.

TABLE 1—ANNUAL REAL GROSS PHYSICAL CAPITAL SECTORAL INVESTMENT  
IN CHILE, 1945–1965: RANGE, MEAN, STANDARD DEVIATION,  
AND PERCENTAGE OF TOTAL INVESTMENT<sup>a</sup>

Real Gross Physical Capital Sectoral Investment					
Sector	Range	Mean	Standard Deviation	As Percentage of Total Gross Real Physical Capital Investment	Real Sectoral Product as Percentage of Total Real Product
Agriculture	25–147	77	38	5	12
Mining	7–140	59	43	4	8
Manufacturing	86–571	327	112	22	20
Transportation	132–606 <sup>b</sup>	338 <sup>b</sup>	151 <sup>b</sup>	20 <sup>b</sup>	7
Utilities	52–274	128	73	8	1
Housing	308–719	519	115	33	7

<sup>a</sup> Data sources are given in the Appendix. All real values are in millions of 1965 Escudos.

<sup>b</sup> For 1950–65 period only.

order of the gross (ignoring depreciation and obsolescence) incremental capital output ratios over the sample, the sectors are utilities, housing, transportation, manufacturing, mining, and agriculture. Such an ordering is consistent with that usually assumed in the development literature (i.e., high incremental capital output ratios in housing and in the social overhead areas). Finally, examination of the last two columns indicates that the six sectors of interest have been much more dominant in total real gross physical capital investment (accounting for 92 percent of the total in the sample) than in total real product (accounting for 55 percent of the total in the sample). An examination of the determinants of real gross physical capital investment in these six sectors, therefore, is almost equivalent to an examination of the determinants of real gross physical capital investment in the Chilean economy but is far from a determination of the total real productive capacity in the Chilean economy because of the large relatively noncapital intensive construction and services (including the government) sectors which are not included in this study (due to lack of data).

In Tables 2 and 3, estimates of the real

gross physical capital sectoral investment relations are presented for the putty-putty model and for the putty-clay model. In Table 4 the implied lag patterns are presented. For each sector the “best” estimate for both the putty-putty and the putty-clay formulations is presented. In addition, several interesting alternative estimates are presented. The information presented in these tables is examined now in respect to ten general observations.

First, on an overall basis, the results suggest that the alternative versions of the underlying putty-putty model are reasonably consistent with Chilean sectoral real gross physical capital investment behavior. The estimates are asymptotically significantly nonzero at the 1 percent level, and the null hypothesis of no serial correlation is not rejected at the same level.<sup>21</sup> The maximum sectoral coefficients of determination suggest that the model is consistent with over half of the variance in the dependent variables in every sector except housing. These results thus seem to support the hypothesis that Chilean

<sup>21</sup> Due to the inclusion of lagged capacity, the Durbin-Watson statistic may be biased towards two although Fisher (1971) speculates that such bias probably is not quantitatively substantial in a similar context.

TABLE 2—ESTIMATES OF ANNUAL REAL GROSS PHYSICAL CAPITAL SECTORAL INVESTMENT FUNCTIONS FOR CHILE, PUTTY-PUTTY MODEL—RELATION 1 (WITH RELATIONS 2, 3a, AND 4 SUBSTITUTED THEREIN), 1945-1965<sup>a</sup>

	Agri- culture A 1	Mining			Manufacturing			Transpor- tation T 1	Utilities			Housing H 1
		Min 1	Min 2	Min 3	Man 1	Man 2	Man 3		U 1	U 2	U 3	
Elasticity of Substitution	0.00	0.00	0.51	1.00	0.00	0.76	1.00	0.00	0.00	0.32	1.00	0.00
Replacement Investment												
1.	183. (3.1)				-31.7 (4.7)	-410. (1.9)	-427. (2.1)	-178. (10.9)	2832. (4.2)	2984. (5.1)	1872. (3.2)	-6.69 (1.4) <sup>b</sup>
2.	-39.1 (3.3)					78.0 (1.8)	81.9 (2.0)		-576. (4.3)	-605. (5.2)	-383. (3.3)	
Neoclassical Term												
1.	-13.9 (1.3) <sup>c</sup>		177. (1.8)	46.4 (1.8)	-50.9 (2.0)	-3.94 (1.8)	-1.88 (1.7)		-1184. (2.8)		305. (1.7)	
2.		-9.82 (2.7)	-86.4 (1.5) <sup>b</sup>	-23.0 (1.6) <sup>b</sup>					174. (2.0)	-307. (3.1)	-365. (1.5) <sup>b</sup>	
3.			10.8 (1.2) <sup>c</sup>	2.85 (1.3) <sup>c</sup>						59.1 (2.6)	127. (1.4) <sup>b</sup>	
4.		.391 (2.3)									-14.2 (1.2) <sup>c</sup>	
Capacity Utilization												
1.		-.303 (1.5) <sup>b</sup>	-1.52 (2.1)	-1.08 (2.0)							-.474 (1.4) <sup>b</sup>	8.65 (3.6)
2.			.0556 (1.6) <sup>b</sup>	.0331 (1.2) <sup>c</sup>								-2.13 (3.8)
Standard Deviation of Relative Price												
1.	.621 (1.3) <sup>c</sup>	6.73 (4.4)	6.15 (2.7)	6.74 (3.1)	5.37 (2.0)	4.20 (1.8)	3.84 (1.7)	-10.1 (2.3)	9.08 (2.4)	12.4 (3.3)	11.1 (2.5)	
2.								2.55 (2.7)	-2.49 (2.6)	-3.25 (3.4)	-2.80 (2.5)	
3.		-.421 (4.6)	-.448 (3.2)	-.467 (3.6)								
Constant	-145. (5.7)	66.9 (9.4)	63.5 (8.6)	62.3 (9.0)	-461. (2.7)	-32.7 (2.6)	-267. (2.4)	-1396. (8.7)	55.6 (1.4) <sup>b</sup>	69.0 (1.9)	64.9 (1.5) <sup>b</sup>	453. (8.4)
$\bar{R}^2$	.83	.53	.42	.49	.49	.58	.59	.89	.70	.76	.69	.43
SE	16.	30.	33.	31.	85.	77.	79.	50.	40.	35.	40.	87.
DW	1.3	1.4	1.4	1.6	1.9	2.2	2.2	2.0	1.4	1.8	1.5	1.5

Note: Elasticity of Substitution is  $\sigma$ , Replacement Investment is  $Qe^{r(\sigma-1)t}$ , Neoclassical Term is  $\Delta(P_Q/P_K)Qe^{r(\sigma-1)t}$ , Capacity Utilization is  $\Delta(Q/Q^C)$ , Standard Deviation of Relative Price is  $\Delta SD(P_Q/P_{GDP})$ .

<sup>a</sup> Data sources are given in the Appendix. The underlying rationales for the models are given in the previous section. The Hall-Sutch method (see fn. 15) is used to allow polynomials of up to the fourth degree (the numbers at the column heads refer to the degree of the relevant polynomials) in the lagged parameters over a four-year period (including the current year). The implied lag patterns are presented in Table 3 below. The absolute value of  $t$ -statistics are given in parentheses beneath the point estimates. Due to the maximum likelihood scanning procedure utilized, all test statistics are asymptotically significantly nonzero at the 5 percent level unless otherwise noted.

<sup>b</sup> Asymptotically significantly nonzero at the 10 percent level.

<sup>c</sup> Asymptotically significantly nonzero at the 15 percent level.

real gross physical capital investment behavior is determined substantially by the same considerations which underlie real gross physical capital investment behavior in more developed economies such as that of the United States.<sup>22</sup> Looking at

the individual sectors on the basis of increasing values of the standard error of estimates, the ordering is agriculture, mining, utilities, transportation, manu-

<sup>22</sup> See the studies mentioned in fn. 3 above.

TABLE 3—ESTIMATES OF ANNUAL REAL GROSS PHYSICAL CAPITAL SECTORAL INVESTMENT  
FUNCTIONS FOR CHILE, PUTTY-CLAY MODEL—RELATION 6, 1945-1965<sup>a</sup>

	Agri- culture A 2	Mining Min 4	Manufacturing		Transpor- tation <sup>e</sup> T 2	Utilities		Housing H 2
			Man 4	Man 5		U 4	U 5	
Elasticity of Substitution ( $\sigma$ )	0.15	0.25	1.00	2.50	0.15	1.00	2.67	0.15
Hicks Neutral Technological Change ( $r$ )	0.00	0.00	0.00	0.005	0.00	0.00	0.00	0.00
First Right-Hand Term in Relation (6)	0.148 (2.0)	0.539 (7.3)	-0.0487 (1.7) <sup>b</sup>	-0.0028 (4.6)	1.97 (4.0)	-0.340 (2.3)	-0.171 (3.5)	0.0492 (0.2) <sup>d</sup>
First Right-Hand Summation in Relation 6 ("Price Term")								
1.	1.89 (0.3) <sup>d</sup>	-0.824 (0.3) <sup>d</sup>	72.2 (5.4)	1.11 (9.4)	-364. (4.5)	-4.01 (2.9)	-0.0499 (3.4)	112. (2.0)
2.			-49.6 (5.4)	-0.767 (9.5)	70.1 (4.4)	0.873 (3.1)	0.0114 (3.5)	-22.7 (2.1)
3.			7.65 (5.6)	0.119 (9.5)				
Second Right-Hand Summation in Relation 6 ("Quantity Terms")								
1.	10.5 (1.3) <sup>e</sup>	23.8 (1.1) <sup>e</sup>		0.280 (1.9)	37.1 (2.5)	31.4 (1.7) <sup>b</sup>	1.69 (1.8) <sup>b</sup>	3.13 (1.0) <sup>e</sup>
2.	-2.23 (1.4) <sup>b</sup>	-4.72 (1.1) <sup>e</sup>	-0.440 (2.3)	-0.0713 (2.3)		-6.56 (1.7) <sup>b</sup>	-0.356 (1.8)	
Standard Deviation of Relative Price								
1.		69.5 (1.4) <sup>b</sup>			-923. (1.2) <sup>e</sup>			
2.					227. (1.5) <sup>b</sup>			
Constant <sup>c</sup>	60.3 (5.2)		442. (9.4)	511. (16.0)	24.3 (5.9)	177. (8.9)	185. (10.9)	483. (11.8)
$\bar{R}^2$	.11	.17	.60	.82	.75	.29	.46	.11
SE	36.	39.	75.	50.	75.	61.	53.	109.
DW	1.0	1.1	1.4	1.6	1.4	0.9	1.4	0.8

Note. See note Table 2.

<sup>a</sup> See Table 2, fn a.<sup>b</sup> Asymptotically significantly nonzero at the 10 percent level.<sup>c</sup> Asymptotically significantly nonzero at the 15 percent level.<sup>d</sup> Not asymptotically significantly nonzero at the 25 percent level.<sup>e</sup> For 1950-65 only.

facturing, and housing. On the basis of the degree of consistency with the variance in the dependent variable the ordering (with the corrected coefficients of determination indicated in parenthesis) is transportation (0.89), agriculture (0.83), utilities (0.76), manufacturing (0.59), mining (0.53), and housing (0.43). With the exception of housing, thus, the putty-putty model is least consistent with the available data on real investment behavior in the two "modern" sectors of mining and manufacturing.<sup>23</sup>

Second, the putty-putty model results are somewhat mixed in respect to the question of what elasticity of substitution between capital and labor is appropriate (i.e., whether the Eisner accelerator assumption, the Jorgenson Cobb-Douglas assumption, or some other assumption is most appropriate). For agriculture, mining, and transportation, the accelerator as-

<sup>23</sup> During the sample Chilean mining was dominated by large international corporations operating primarily in copper. For details see Markos Mamalakis (1967), Clark Reynolds, and Mario Vera Valenzuela.

TABLE 4—LAG PATTERNS FOR ESTIMATES IN TABLES 2 AND 3

Estima- tion Number	Investment Determinant <sup>a</sup>	-0	-1	Lag Pattern -2	-3	Σ
A 1	Replacement	-.106	.0378	.103	.0908	.126
	$\Delta K^D$ : neoclassical	.0554	.0416	.0277	.0139	.139
	SD relative price	-2.49	-1.86	-1.24	-.621	-6.21
A 2	Putty-Clay: price	-.755	-.566	-.378	-.189	-1.89
	quantity	-.621	.204	.582	.514	.679
Min 1	$\Delta K^D$ : neoclassical	.048	.044	.025	-.002	.115
	capacity utilization	121.	90.9	60.6	30.3	303.
	SD relative price	0.0	6.3	10.1	8.9	25.3
Min 2	$\Delta K^D$ : neoclassical	-0.175	.0836	.0769	.0274	.170
	capacity utilization	253.	106.	-7.	-54.	299.
	SD relative price	4.08	9.78	12.8	10.4	37.1
Min 3	$\Delta K^D$ : neoclassical	.0001	.0263	.0237	.0092	.0594
	capacity utilization	221.	116.	31.	-14.	355.
	SD relative price	2.91	9.19	12.7	10.5	35.3
Min 4	Putty-Clay: price	.330	.247	.165	.082	.824
	quantity	-1.96	-.053	.909	.926	-.178
	SD relative price	-2.78	-2.08	-1.39	-.695	-6.94
Man 1	Replacement	.127	.095	.063	.032	.317
	$\Delta K^D$ : neoclassical	.204	.153	.102	.051	.509
	SD relative price	-21.5	-16.1	-10.7	-5.4	-53.7
Man 2	Replacement	.394	.0610	-.115	-.136	.203
	$\Delta K^D$ : neoclassical	.0158	.0118	.00789	.00394	.0394
	SD relative price	-16.8	-12.6	-8.40	-4.20	-42.0
Man 3	Replacement	.396	.051	-.129	-.147	.171
	$\Delta K^D$ : neoclassical	.00751	.00563	.00376	.00188	.0188
	SD relative price	-15.4	-11.5	-7.69	-3.85	-38.5
Man 4	Putty-Clay: price	1.46	4.49	2.20	-0.83	7.33
	quantity	.703	.659	.527	.308	2.20
Man 5	Putty-Clay: price	.0226	.0691	.0333	-.0135	0.112
	quantity	.0022	.0230	.0296	.0219	0.768
T 1	Replacement	.710	.533	.355	.178	1.78
	$\Delta K^D$ : SD relative price	-.400	-7.90	-10.4	-7.70	-26.4
T 2	Putty-Clay: price	33.4	4.02	-11.3	-12.7	13.4
	quantity	-14.9	-11.1	-7.4	-3.7	-37.1
	SD relative price	6.0	-63.6	-87.8	-66.6	-212.
U 1	Replacement	-2.11	.014	1.25	1.20	.480
	$\Delta K^D$ : neoclassical	1.95	.94	.28	-.03	3.14
	SD relative price	3.5	10.1	11.7	8.4	33.7
U 2	Replacement	-2.25	.124	1.29	1.25	.415
	$\Delta K^D$ : neoclassical	1.13	.886	.378	-.0355	2.36
	SD relative price	2.60	11.7	14.3	10.4	39.0
U 3	Replacement	-1.36	.130	.853	.810	.435
	$\Delta K^D$ : neoclassical	.134	.186	.071	.039	.430
	capacity utilization	190.	142.	94.8	47.4	474.
	SD relative price	.368	8.67	11.4	8.49	28.9
U 4	Putty-Clay: price	.206	-.108	-.246	-.211	-.359
	quantity	-2.06	.43	1.60	1.45	1.42
U 5	Putty-Clay: price	.0017	-.0021	-.0037	-.0030	-.007
	quantity	-.106	.027	.089	.080	.090
H 1	Replacement	.0268	.0201	.0134	.0067	.0667
	$\Delta K^D$ : capacity utilization	-42.6	606.	829.	623.	2015.
H 2	Putty-Clay: price	-8.31	-.58	4.93	4.73	1.92
	quantity	-1.25	-.94	-.63	-.31	-3.13

<sup>a</sup> Investment Determinants refer to column heads in Tables 2 and 3.

sumption leads to the most consistency with the data. For manufacturing, the results are insensitive between the estimated value from Behrman (1972b) and a value of one, but are less satisfactory when the assumed value is zero. For utilities, the value estimated by Behrman (1972b) for this elasticity leads to greater consistency with the variance in the dependent variable than does a value either of zero or of one. In other words, if one always made the accelerator assumption of a zero elasticity of substitution, one would forego some degree of consistency with the variance in the dependent variable in manufacturing and utilities. If one always made the assumption of an elasticity of substitution as estimated by Behrman (1972b),<sup>24</sup> one would forego some degree of consistency in agriculture and mining. If one always made the assumption of an elasticity of substitution of one, one would forego some consistency in agriculture, mining, transportation, utilities, and housing. Thus, although a particular a priori assumption about the value of the elasticity of substitution between capital and labor in the putty-putty formulation *may* not make much difference, in some cases the results do seem quite sensitive to the value of this elasticity and it would seem somewhat dangerous to limit one's consideration to a particular a priori value such as zero or one.<sup>25</sup>

Third, in respect to real replacement investment in the putty-putty model, one

or the other of the two alternatives mentioned above in Section I dominates in each cell. For mining alone, the first or complete putty-putty alternative dominates. For this sector, the estimated relation is less satisfactory if a distributed lag of capacity levels (corrected for technological changes) is included, with the apparent implication that capital is relatively malleable (or depreciation and obsolescence are relatively limited) in mining. For the other five sectors, the second or partial putty-putty alternative dominates. The significance and size of the estimated coefficients for the lag pattern of weights of present and past capacity levels (corrected for technological changes) indicate that relatively inflexible replacement requirements determine a substantial portion of real gross physical capital investment. Some empirical support, therefore, is provided for Jorgenson and his collaborators'<sup>26</sup> representation of real replacement physical capital needs by the lagged level of existing real physical capital stock and thus for the partial as opposed to the complete putty-putty model.

Fourth, on an overall basis, the results suggest that the putty-clay model is less satisfactory than the putty-putty model in being consistent with Chilean sectoral real gross physical capital investment behavior. Problems of positive serial correlation apparently are greater in the putty-clay formulation and only in manufacturing is the standard error of estimate lower and the coefficient of determination higher in the case of the putty-clay model than in the best putty-putty estimate. Among the putty-clay estimates, moreover, for all sectors except manufacturing the sum of the weights for either the price term or for the quantity term is nonpositive (although

<sup>24</sup> Including a zero value for housing for which Behrman (1972b) presents no estimates because of the lack of labor and wage data for this sector.

<sup>25</sup> Of course, these comments do not take into account the criticisms of time-series estimates of the elasticity of substitution emphasized by Hall and Jorgenson (1969). However, the only information available about cross-section estimates of the elasticity of substitution in Chile is some work in process for manufacturing at the Instituto de Economía which also suggests that this elasticity in manufacturing is significantly different from both zero and one.

<sup>26</sup> See Jorgenson (1963), p. 254, Jorgenson and Siebert, p. 682, and Jorgenson and Stephenson (1967a), p. 192-212.

in several cases this sum probably is not significantly nonzero). For two sectors (including the otherwise relatively satisfactory case of manufacturing), furthermore, the first term in relation (6) is asymptotically significantly negative, a result which also is not easily interpretable.<sup>27</sup>

Fifth, in each sector for the putty-clay model, scanning over the elasticity of substitution between capital and labor led to maximizing values outside of the initial ranges indicated above in Section I.<sup>28</sup> For manufacturing and utilities, the estimated elasticities of substitution are quite high (2.50 and 2.67, respectively).<sup>29</sup> For the other four sectors, the estimates are quite low (at the lower limit of 0.15 except in the case of 0.25 for mining). The best putty-putty estimates for all four of the sectors in this latter group imply a simple accelerator model, which is consistent with the low values for the elasticity of substitution obtained in the putty-clay version. However, the putty-putty formulation is much more consistent with variations in real gross sectoral investment in all four sectors than is the putty-clay model. Examination of the estimates for agriculture, transportation, and housing suggests that the greater consistency of the putty-putty model may be due to the real replacement investment formulation in the partial putty-putty alterna-

tive. Except for this real replacement term and the Taylor series approximation, the basic putty-clay model reduces to the basic putty-putty model as the elasticity of substitution between capital and labor approaches zero. Examination of the estimates for mining, however, suggests that the real replacement investment formulation in the partial putty-putty model is not the only explanation for the difference between the degree of consistency with the variance in the dependent variable in the two models. For mining the complete putty-putty alternative of the putty-putty model (without explicit inclusion of real replacement investment) is preferable to the partial putty-putty alternative. Perhaps the difference in part lies in inadequacies in the Taylor series approximation for the putty-clay model,<sup>30</sup> in which case the comparison between the models in the previous observation may be unfairly biased against the putty-clay formulation.

Sixth, in all sectors for the putty-clay model the estimates are not very sensitive to the value of the Hicks neutral exponential technological change parameter. For the best estimate of the elasticity of substitution, however, no Hicks neutral exponential technological change is implied (except for manufacturing, in which case a low rate of 0.005 results). Thus, substantially lower rates of Hicks neutral exponential technological change are implied for mining, manufacturing, and transportation than are reported in Behrman (1972b), but given the qualifications in the previous two observations, one can not have much confidence in the reliability of these estimates from the putty-clay model.

<sup>27</sup> In light of the unsatisfactory aspects of the putty-clay model estimates, the fifth and sixth comment may be of quite limited significance. Also, in these and in subsequent comments emphasis is placed on the putty-putty results.

<sup>28</sup> The lower limit on such scanning was 0.15 since such a value is quite close to zero (at which value the model reduces to an accelerator model). No upper limit was established. Instead, the value of the elasticity of substitution was increased as long as such increases augmented the degree of consistency with the variance in the dependent variable.

<sup>29</sup> For these two sectors the best estimate of the putty-putty model also implied relatively high elasticities of substitution. Note, furthermore, that only for these two sectors is (coincidentally?) the estimate of the first term in relation (6) asymptotically significantly negative.

<sup>30</sup> This possibility is not explored further in this study due to the considerable costs which such exploration would require, but it should be kept in mind in consideration of the implications of the estimates for any comparison of the two basic underlying real investment models.

Seventh, considering all the estimates, significant responses to relative prices in Chilean sectoral real physical capital investment decisions do not seem to be widespread. Such responses apparently are important, however, in manufacturing, which generally is considered to be a key sector in the developing countries. These responses also may be important in utilities and, perhaps, in mining. Thus, apparently some support is given to part of the Latin American "structuralists" position in that the modern sectors are relatively price responsive in respect to capacity expansion, while agriculture is not.<sup>31</sup>

Eighth, the capacity utilization term apparently is a significant determinant of the desired real physical capital stock only in the putty-putty formulation in mining and housing and possibly in utilities but not otherwise. Real physical capital investment in these sectors thus apparently is responsive to policies (for example, fiscal and monetary policies which affect aggregate demand) which affect the utilization of existing capacity.<sup>32</sup> Perhaps this capacity utilization variable is not significant in more sectors because actual utilization rates of *rated* capacity are so low that fluctuations therein are irrelevant in decisions relating to the expansion of capacity through real gross physical capital investment.<sup>33</sup> Another possibility (which was not explored in this study) is that the utilization rate (or expectations thereof) affects not the level of desired real capital physical stock but the path of adjustment of the actual real capital physical stock to the desired level.

<sup>31</sup> See Enrique Sierra, pp. 34-50 for a summary of the major components of the Latin American structuralist position.

<sup>32</sup> The determinants of the utilization rates in these sectors are relative prices, the level of aggregate demand, and the availability of intermediate imports. See Behrman (1973a).

<sup>33</sup> Phillips has suggested that this might widely be the case at least in the manufacturing sectors of developing economies.

Ninth, the standard deviation of the product price relative to the gross domestic product deflator term apparently is a significant determinant of the desired real physical capital stock in every sector except housing in the putty-putty formulation, although the evidence of a widespread response to this variable is substantially less in the putty-clay formulation. For the putty-putty model for agriculture, manufacturing, and transportation and for the putty-clay model for mining and transportation, the estimates of the coefficients of the variables related to this determinant indicate risk aversion, which suggests that greater stability in relative price movements would increase real net physical capital investment. To the extent that instability in relative prices is related to differing patterns of adjustment to changes in the rates of change of the nominal money supply and of the exchange rate,<sup>34</sup> lessening relative price fluctuations by lessening inflation may serve to increase the real net physical capital investment in these sectors as well as serve a number of other announced goals of the government. For the putty-putty model for the foreign controlled mining sector and for utilities, however, the estimates of the coefficients of the relevant variables indicate risk preference with the opposite implications.

Tenth, for the polynomials underlying the estimated lag patterns (which are presented in Table 4) generally one or two polynomial terms of degree one, two, three, or four have significantly nonzero coefficient estimates.<sup>35</sup> In the cases in

<sup>34</sup> The results in Behrman (1973c) suggest that such factors may underlie a considerable portion of the variance in relative prices.

<sup>35</sup> But in Min 2 and Min 3 three polynomial terms for the neoclassical term have significantly nonzero coefficients, in U 3 four polynomial terms have significantly nonzero coefficients for the neoclassical term, and in Man 4 and Man 5 three polynomial terms for the putty-clay price impact have significantly nonzero coefficients.

which only the coefficient estimate for one polynomial is significantly nonzero, the lag pattern, of course, has an (absolute) maximum for the current period and declining (absolute) values thereafter. In the thirty-four cases in which more than one polynomial term has a significantly nonzero coefficient estimate, the maximum weight occurs 11 times in the current year, 5 times with a lag of one year, 17 times with a lag of two years, and 2 times with a lag of three years. That over half of the lag patterns apparently are better represented by more than one polynomial and that the maximum weight often is not for the current period suggest that the quite frequently used linear or geometric representations of lag patterns would be less adequate than the Hall-Sutch procedure. For the one sector for which the putty-clay estimates are most satisfactory, manufacturing, the estimated lag patterns imply quick relative adjustment to the price terms in comparison to the adjustment to the quantity terms in contradiction to the implications of the putty-clay hypothesis.<sup>36</sup> Finally, in respect to sign reversals in the lag patterns, in a number of cases (i.e., A 1, A 2, Min 2, Min 4, T 2, U 1-5, and

H 2) the implied weight for the current period for at least one term is negative. One possible implication of such patterns is that the endpoint of the lag pattern should be pinned to a value of zero at time zero.<sup>37</sup> In a number of other cases (i.e., Min 1-3, Man 2-3, U 1-2, and U 4-5) the implied weights for the period lagged three years (and in some cases, two years) are negative. Such a pattern of weights might reflect overshooting to correct for some initial overadjustment<sup>38</sup> or might reflect the use of too long a lag period.

### III. Conclusions

Although questions have been raised about the widely assumed dominance of real physical capital stock investment in the economic development process, the importance of real physical capital stock as one of the major factors in development still is widely accepted. The question of what determines real physical capital investment behavior in developing countries, therefore, is a very important one. Estimates of real physical capital stock investment functions on the sectoral level for the postwar Chilean experience suggest that this experience has been substantially consistent with the putty-putty type of investment behavior models which have been used for the United States. Evidence has been presented of the widespread significance in Chilean real investment behavior of real replacement needs, of neoclassical considerations based on a CES production function, of higher moments of the subjective probability function as represented by the standard deviation of prices, and, to a lesser extent, of capacity utilization considerations. Therefore, government policies can induce in-

<sup>36</sup> The statement in the text need be qualified only in respect to the current adjustment in estimation number Man 4. The accumulated responses as a percentage of the total responses for putty-clay manufacturing estimates are as follows:

	Lag (in percentage)			
	-0	-1	-2	-3
Man 4 Putty-clay				
price	20	81	111	100
quantity	32	62	84	100
Man 5 Putty-clay				
price	20	82	112	100
quantity	03	33	72	00

The relatively quick adjustment to the price terms may reflect a relatively high price consciousness among Chileans due to their long inflationary history (see fn. 6 above). Given the overall weaknesses of the putty-clay estimates, however, such an interpretation has very limited support from this study.

<sup>37</sup> Distributed lags over alternative periods were not explored in this study.

<sup>38</sup> For examples of such overshooting in other contexts in the Chilean economy, see Behrman (1973c) and Behrman and Jorge García.

creased real physical capital investment, particularly in the key sector of manufacturing, by increasing the price of output relative to the price of capital services through reductions in the cost of capital or in the effective direct business tax rate, through increases in depreciation allowances or investment tax credits, and through foreign trade policy. Government fiscal and monetary policies which increase the degree of capacity utilization and reduce uncertainty also may result in increased real physical capital investment in various sectors.

In addition to the evidence about real investment behavior in a particular developing economy, the results have several important implications for the general examination of investment behavior. For one, the lag patterns suggest the need for more flexibility than allowed by geometric lags. For another, the results suggest that considerations related solely to expectations might usefully be supplemented at least by consideration of second moments. Moreover, the results suggest that the estimates *may* be sensitive to the particular assumption made about the elasticity of substitution and that a priori specifications of any particular value for this elasticity (such as a value of zero for a pure accelerator model or a value of one for Jorgenson's neoclassical Cobb-Douglas model) should not be accepted unquestioned. Finally, at least for the data under examination, the estimates suggest that the partial putty-putty formulation of the real gross physical capital investment decision is superior to the approximation to the putty-clay model which was explored. This apparent superiority may well reflect the treatment of real replacement investment, however, much more than the difference between putty-putty and putty-clay treatment of real net investment.

## APPENDIX

### *Data Sources*

Nominal values for sectoral and total value-added, total nominal direct business taxes, total nominal physical capital investment in housing and for the economy, and gross domestic product deflator: 1940-57 from Corporación de Fomento (CORFO), Dirección de Planificación, Departamento de Investigaciones Económicas, "Cuentas Nacionales de Chile, 1940-1962" (provisional revised figures, Santiago: CORFO, June 1964, mimeo); 1964-1965 from ODEPLAN, Presidencia de la Republica, "Cuentas Nacionales de Chile, 1964-1965" (estimated figures, Santiago, ODEPLAN, July 1966, mimeo).

Nominal values for gross domestic product were deflated by the sectoral price indices (see below) to obtain series for real sectoral gross domestic products. A trend-through-the-peaks procedure (see Klein and Summers) then was used to determine sectoral values of the capacity of real gross domestic product.

### *Sectoral Prices and Deflators:*

1. Agriculture: agricultural component of the index of wholesale prices from various issues of *Boletín Mensual del Banco Central*.
2. Mining: constructed in ODEPLAN on basis primarily of the index of the unit value of mineral exports (*Balanza de Pagos*, Banco Central), and the exchange rate used for the national accounts (same source as for sectoral nominal value-added).
3. Manufacturing: same as agriculture.
4. Transportation: 1940-61, on the basis of weights for value-added in 1960 (from División Transporte, ODEPLAN), weighted average of (a) index of average railroad tariff for passengers and cargo (*Memorias Anuales de FF.CC. de Estado*) and (b) index of urban autobus fares (Dirección de Estadística y Censos). 1960-65—Index of the average tariff for transportation (División Transporte, ODEPLAN).
5. Utilities based on index of average

tariff of KWH (*Memorias Anuales de CHILECTRA*).

6. Housing Services: 1940–1957, housing services component of index of cost of living (*Boletín Mensual del Banco Central*); 1958–1968, housing services component of index of consumer prices (*Boletín Mensual del Banco Central*).

Bank interest rates: various issues of *Boletín Mensual del Banco Central*.

Sectoral depreciation estimates: División de Programación Global, ODEPLAN (unpublished).

Sectoral investment goods price index: based on sectoral prices (see above) and the weights given in Meza.

Sectoral real gross physical capital investment: División de Precios, ODEPLAN (unpublished).

Most of the national accounts data also are presented and discussed in the comprehensive study of Chilean historical data by Mamalakis.

#### REFERENCES

- I. Adelman and F. T. Sparrow, "Experiments with Linear and Piece-Wise Linear Dynamic Programming Models," in I. Adelman and E. Thorbecke, eds., *The Theory and Design of Economic Development*, Baltimore 1966, 291–316.
- K. J. Arrow, H. B. Chenery, B. S. Minhas, and R. M. Solow, "Capital-Labor Substitution and Economic Efficiency," *Rev. Econ. Statist.*, Aug. 1961, 45, 225–50.
- J. R. Behrman, (1973a) "Cyclical Sectoral Capacity Utilization in a Developing Economy," in R. E. Eckhaus and P. N. Rosenstein-Rodan, eds., *Analysis of Development Problems: Studies of the Chilean Economy*, Amsterdam 1973 forthcoming, Ch. 11.
- , (1973b) "Aggregative Market Responses in Developing Agriculture: The Post-war Chilean Experience," in R. Eckhaus and P. N. Rosenstein-Rodan, eds., *Analysis of Development Problems: Studies of the Chilean Economy*, Amsterdam 1973 forthcoming, Ch. 9.
- , (1973c) "Price Determination in an Inflationary Economy: The Dynamics of Chilean Inflation Revisited," in R. Eckhaus and P. N. Rosenstein-Rodan, eds., *Analysis of Development Problems: Studies of the Chilean Economy*, Amsterdam 1973 forthcoming, Ch. 15.
- , (1972a) "Review Article: UNCTAD Secretariat, *Trade Prospects and Capital Needs of Developing Countries*," *Int. Econ. Rev.*, Oct. 1971, 12, 519–25.
- , (1972b) "Sectoral Elasticities of Substitution between Capital and Labor in a Developing Economy: Time Series Analysis in the Case of Postwar Chile," *Econometrica*, Mar. 1972, 40, forthcoming.
- , (1972c) "Short Run Flexibility in a Developing Economy," *J. Polit. Econ.* Mar./Apr. 1972, 80, 292–313.
- and J. García M., "A Study of Quarterly Nominal Wage Change Determination in an Inflationary Developing Economy," in R. Eckhaus and P. N. Rosenstein-Rodan, eds., *Analysis of Development Problems: Studies of the Chilean Economy*, Amsterdam 1973, forthcoming, Ch. 16.
- C. W. Bischoff, "The Effect of Alternative Lag Distributions," in G. Fromm, ed., *Tax Incentives and Capital Spending*, Washington 1971, 61–130.
- , "Hypothesis Testing and the Demand for Capital Goods," *Rev. Econ. Statist.*, Aug. 1969, 51, 354–68.
- M. Brown, *On the Theory and Measurement of Technological Change*, Cambridge 1966.
- , *The Theory and Empirical Analysis of Production*, Nat. Bur. Econ. Res. Stud. in *Income and Wealth*, Vol. 31, New York 1967.
- H. Chenery and A. Strout, "Foreign Assistance and Economic Development," *Amer. Econ. Rev.*, Sept. 1966, 56, 679–733.
- R. M. Coen, "The Effect of Cash Flow on the Speed of Adjustment," in G. Fromm, ed., *Tax Incentives and Capital Spending*, Washington 1971, 131–96.
- , "Effects of Tax Policy on Investment in Manufacturing," *Amer. Econ. Rev. Proc.*, May 1968, 58, 200–11.
- , "Tax Policy and Investment Behav-

- ior: Comment," *Amer. Econ. Rev.*, June 1969, 59, 370-79.
- R. Eisner, "A Distributed Lag Investment Function," *Econometrica*, Jan. 1960, 28, 1-29.
- , "A Permanent Income Theory for Investment: Some Empirical Explorations," *Amer. Econ. Rev.*, June 1967, 57, 363-90.
- , (1969a) "Investment and the Frustrations of Econometricians," *Amer. Econ. Rev. Proc.*, May 1969, 59, 50-64.
- , (1969b) "Tax Policy and Investment Behavior: Comment," *Amer. Econ. Rev.*, June 1969, 59, 379-88.
- and M. I. Nadiri, "Investment Behavior and Neo-Classical Theory," *Rev. Econ. Statist.*, Aug. 1968, 50, 368-82.
- and R. H. Strotz, "Determinants of Business Investment," in *Impacts of Monetary Policy, Commission on Money and Credit*, Englewood Cliffs 1963, 59-337.
- F. M. Fisher, "Discussion," in G. Fromm, ed., *Tax Incentives and Capital Spending*, Washington 1971, 243-55.
- , "The Existence of Aggregate Production Functions," *Econometrica*, Oct. 1969, 37, 553-77.
- P. Gregory, *Industrial Wages in Chile*, Ithaca 1967.
- R. E. Hall and D. W. Jorgenson, "Application of the Theory of Optimum Capital Accumulation," in G. Fromm, ed., *Tax Incentives and Capital Spending*, Washington 1971, 9-60.
- and ———, "Tax Policy and Investment Behavior," *Amer. Econ. Rev.*, June 1967, 57, 391-414.
- and ———, "Tax Policy and Investment Behavior: Reply and Further Results," *Amer. Econ. Rev.*, June 1969, 59, 388-401.
- R. Hall and R. C. Sutch, "A Flexible Infinite Distributed Lag," Institute of Business and Economics Research Center for Research in Management Science, Univ. California, Berkeley 1969.
- A. Harberger, "The Dynamics of Inflation in Chile," in Carl Christ, ed., *Measurement in Economics: Studies in Mathematical Economics in Memory of Yehuda Grunfeld*, Stanford 1963, 219-50.
- and M. Selowsky, "Key Factors in the Economic Growth of Chile: Analysis of the Sources of Past Growth and of Prospects for 1965-1970," paper presented at conference on "The Next Decade of Latin American Economic Development," Cornell Univ., Apr. 20-22, 1966.
- D. W. Jorgenson, "Anticipation and Investment Behavior," in J. S. Duesenberry et al., eds., *The Brookings Quarterly Econometric Model of the United States*, Chicago 1965, 35-92.
- , "Capital Theory and Investment Behavior," *Amer. Econ. Rev. Proc.*, May 1963, 53, 247-59.
- and C. D. Siebert, "A Comparison of Alternative Theories of Corporate Investment Behavior," *Amer. Econ. Rev.*, Sept. 1968, 58, 681-712.
- and J. A. Stephenson, (1967a) "Investment Behavior in U.S. Manufacturing, 1947-60," *Econometrica*, Apr. 1967, 35, 169-220.
- and ———, (1967b) "The Time Structure of Investment Behavior in U.S. Manufacturing, 1947-60," *Rev. Econ. Statist.*, Feb. 1967, 49, 16-27.
- L. R. Klein and R. Summers, *The Wharton Index of Capacity Utilization*, Philadelphia 1966.
- and P. Taubman, "Estimating Effects within a Complete Econometric Model," in G. Fromm, ed., *Tax Incentives and Capital Spending*, Washington 1971, 197-242.
- M. Mamalakis, "The American Copper Companies and the Chilean Government, 1920-1967: Profile of an Export Sector," Yale Economic Growth Center Disc. Pap. No. 37, Yale University 1967.
- , "Historical Statistics of Chile" (4 Vols.), New Haven.
- W. Meza, "Inversión Geográfica Bruta en Capital Fijo Por Sectores de Destino, Período 1962-1966," Oficina de Planificación Nacional, División de Programación Financiera, mimeo., Santiago 1968.
- A. Phillips, "An Appraisal of Measures of Ca-

- capacity," *Amer. Econ. Rev. Proc.*, May 1963, 53, 275-92.
- , "Measuring Industrial Capacity in Less Developed Countries," Dept. Econ. Disc. Pap. No. 110, Univ. Pennsylvania 1969.
- J. Ramos, "Políticas de Remuneraciones en Chile," Universidad de Chile, Instituto de Economía, Santiago, mimeo., Sept. 1968.
- C. Reynolds, "Development Problems of an Export Economy: The Case of Chile and Copper," in M. Mamalakis and C. Reynolds, *Essays on the Chilean Economy*, Homewood 1965, 201-398.
- E. Sierra, *Tres Ensayos de Establización de Chile: Las Políticas Aplicados en el Decenio 1956-66*, Santiago 1969.
- R. Summers, "Further Results in the Measurement of Capacity Utilization," in *Proc. Bus. Econ. Sec. Amer. Statist. Ass.*, 1968, 25-34.
- M. V. Valenzuela, *La Política Económica del Cobre en Chile*, Universidad de Chile, Santiago 1961.
- United Nations, Conference on Trade and Development, *Trade Prospects and Capital Needs of Developing Countries*, New York 1968.

# Capital Deepening Response in an Economy with Heterogeneous Capital Goods

By EDWIN BURMEISTER AND STEPHEN J. TURNOVSKY\*

Capital deepening is an important concept in traditional capital theory. In a one-sector model it has an unambiguous definition, describing an increase in the physical capital-labor ratio. Moreover, the one-sector model always exhibits capital deepening when one compares a steady-state equilibrium having a high interest or profit rate with one having a low rate. We thus define *capital deepening response* in the one-sector model as an equilibrium increase in the physical capital-labor ratio in response to a decrease in the steady-state interest or profit rate. An economy which exhibits such capital deepening response for every admissible interest or profit rate will be termed *regular*. Furthermore, it is a fundamental result that steady-state consumption behavior is not paradoxical (as defined below) if, and only if, the economy is regular.

When the model includes heterogeneous capital goods, there is in general no unambiguous physical definition of capital deepening. If some capital-labor ratios rise while others fall, we cannot say unambiguously that there is capital deepening in any physical sense. Nevertheless, it is clearly desirable to have a definition of capital deepening response and a definition of a regular economy which preserve

the basic properties of the one-sector model.

The "Cambridge (U.K.) approach" (see, for example, L. L. Pasinetti (1969, 1970) and Joan Robinson) is to define capital deepening across alternative steady-state equilibria in terms of the change in the *value* of the per capita capital stock. However, as we demonstrate, this definition does not generalize the results of the one-sector model. We propose an alternative definition which is not only consistent with the conventional concept of capital deepening response in one-capital good models but also is intimately related to the phenomenon of paradoxical consumption behavior. In particular, there is no paradoxical consumption behavior if, and only if, the economy exhibits capital deepening response, as we define it, at every admissible interest or profit rate and which, therefore, is regular in our terminology. This relationship points out the relevance of our concepts since it is consumption behavior and not the behavior of capital values that should be our ultimate economic concern.

For expositional purposes we limit ourselves to one consumption good or possibly a fixed basket of consumption goods. Otherwise, the change in the consumption basket mix due to relative price changes would have to be considered, and we wish to avoid this issue. Thus, we are concentrating on problems caused by the existence of heterogeneous capital goods rather than heterogeneous consumption goods.

Throughout this paper we are concerned

\* Professors of economics, University of Pennsylvania, Australian National University, respectively. Financial support from the National Science Foundation is gratefully acknowledged by Burmeister. We have also benefited from conversations with Paul A. Samuelson and comments by Stephen A. Ross.

only with a comparison of steady-state equilibria, and we take the profit or interest rate, denoted by  $r$ , as exogenous. Thus we are *not* considering a fully determined general equilibrium model. One advantage of this approach is that our results do not depend on the exact relationships that are needed to form a complete model in which the equilibrium value of  $r$  is determined simultaneously as one of many endogenous variables. In particular, our approach does not require the specification of any savings-investment behavior, and thus we avoid unnecessary complexities and obtain quite general results. Of course, as a consequence we cannot study many important economic questions about dynamics; for example, we cannot analyze the tradeoff between future and present consumption. And nothing we say, either for the joint or no-joint production case, is meant to imply that it is in any sense "optimal" or "efficient" for an economy to remain in any particular steady-state equilibrium except under special circumstances.

Finally, for expositional convenience we omit discussion of linear Leontief technologies; but as we shall indicate at the appropriate place below, our results are easily generalized to the nondifferentiable case.<sup>1</sup> It is well known from the reswitching controversy that paradoxical consumption behavior in linear technologies does indeed occur (see, for example, Michael Bruno, Burmeister, and Eytan Sheshinski; Paul Samuelson (1966b)). One of the objectives of this paper is to derive a necessary and sufficient condition which excludes such behavior in both neoclassical and linear models.<sup>2</sup>

<sup>1</sup> See Christian von Weizsäcker, pp. 60–66.

<sup>2</sup> While we have not seen a numerical example of a neoclassical technology where paradoxical consumption behavior does occur, the fact that it exists in linear models which can be approximated as closely as desired by smooth production functions strongly suggests that it can occur in the neoclassical case as well.

### I. Capital Deepening in the One-Sector Model

Consider the standard neoclassical one-sector model in which there is a single output, part of which is consumed, the rest of which is accumulated as capital for further production. Gross output  $Y$  is given by the neoclassical production function, satisfying the usual regularity conditions,<sup>3</sup>

$$(1) \quad Y = F(K, L)$$

where  $K$  and  $L$  denote the capital and labor inputs, respectively. Because of the homogeneity of  $F$ , this production function can be written in the per capita form

$$(2) \quad y = f(k)$$

where  $y = Y/L$ ,  $k = K/L$ ,  $f'(k) > 0$ ,  $f''(k) < 0$ .<sup>4</sup> Assuming that capital depreciates at a constant rate  $\delta$ , we know that  $\dot{K} = I - \delta K$  where  $I$  is gross investment and the dot denotes the time derivative. Thus, from the national income identity  $Y = C + I$  (where  $C$  is consumption) we deduce that per capita consumption,  $c = C/L$ , must be given by

$$(3) \quad c = f(k) - \dot{k} - (g + \delta)k$$

(where  $g$  is the rate of growth of labor). Hence, the steady-state per capita consumption is

$$(4) \quad c = f(k) - (g + \delta)k$$

As is well known, under competitive conditions the steady-state capital-labor  $k$  is the solution to

$$(5) \quad f'(k) = r + \delta$$

<sup>3</sup>  $F_i$  denotes the partial derivative of  $F$  with respect to its  $i$ th argument ( $i = 1, 2$ ) and is assumed positive. The matrix  $[F_{ij}]$ , which is the Hessian of  $F$ , is negative semidefinite. Also,  $F$  is assumed to be homogeneous of degree one and twice continuously differentiable.

<sup>4</sup> Here primes denote derivatives with respect to the single argument  $k$ . Later in the paper, as will be clearly indicated, primes will denote the transpose operator and, when the context makes our usage clear, differentiation with respect to  $r$ .

where  $r$  denotes the rate of interest, assumed given exogenously.<sup>5</sup> Since  $f''(k) < 0$ ,  $k$  can be solved uniquely in terms of  $r$ . The question of capital deepening response involves the response of  $k$  to  $r$ ; as we see, the answer to this question is very simple in the present one-sector model. Differentiating (5) with respect to  $r$ , we find

$$(6) \quad \frac{dk}{dr} = \frac{1}{f''(k)} < 0$$

Hence, we deduce unequivocally that a decrease in the interest rate  $r$  will cause an increase in the steady-state capital-labor ratio  $k$ ; such an economy exhibits *capital deepening response* for all admissible values of  $r$  and therefore is regular.

At the same time, since our ultimate interest is in consumption behavior, we wish to determine how the corresponding steady-state per capita consumption responds to  $r$ . Again, the solution is readily determined. Differentiating (4) with respect to  $r$  and using the competitive pricing conditions (5), we obtain

$$(7) \quad \frac{dc}{dr} = (r - g) \frac{dk}{dr}$$

implying that  $c$  has a relative maximum at  $r=g$ , which, of course, is the Golden Rule. Furthermore, by definition *the model shows no paradoxical consumption behavior* if, and only if,

$$(8) \quad \operatorname{sgn} \frac{dc}{dr} = \operatorname{sgn} (g - r)$$

for all admissible values of  $r$ . That is, there is no paradoxical consumption be-

havior if per capita consumption increases for all  $r < g$ , and decreases for all  $r > g$ . Conversely, we say that the model shows paradoxical consumption behavior if there is some range of interest rates in which per capita consumption increases while the gap  $|r - g|$  increases. However, since in the one-sector model  $dk/dr < 0$ , it follows that the condition (8) is indeed satisfied so that  $c$ , expressed as a function of  $r$ , has a unique maximum at  $r=g$ .

Thus, the simple one-sector neoclassical model is regular; the capital-labor ratio always varies inversely with the rate of interest. Furthermore, as a consequence of this property, we find that per capita consumption is "well-behaved," being free of the above paradox. It also can be shown that exactly the same conclusions remain valid for the neoclassical two-sector model in which there again is only one type of physical capital good; see Burmeister.

## II. The Heterogeneous Capital Good Model: The No-Joint Production Case

In this section we shall give a definition of capital deepening response which generalizes the property of the one-sector definition to many capital goods. We shall assume that the economy's production possibilities can be summarized by a *production possibility frontier (PPF)* of the form

$$(9) \quad c = T(y_1, \dots, y_n; k_1, \dots, k_n)$$

where

$k_i$  = per capita quantity of the  $i$ th capital good ( $i = 1, \dots, n$ )

$y_i$  = per capita output of the  $i$ th capital good.

Analogous to (3) we have the accumulation equations

$$(3') \quad \dot{k}_i = y_i - (g + \delta_i)k_i \quad i = 1, \dots, n$$

where  $\delta_i$  denotes the exponential depreciation rate for the  $i$ th capital good (which

<sup>5</sup> Throughout this paper we will ignore corner or boundary solutions. Thus we must assume that every exogenously given value of  $r$  belongs to an open interval for which (5) has a solution with a positive value of  $k$ . We also assume that  $r=g$  belongs to this interval of admissible  $r$ 's which we denote by  $R$ . Analogous assumptions are needed in models with many capital goods; see fn. 6 below.

for convenience is assumed to be independent of good  $j$ ). Hence, in a steady state we have

$$(3'') \quad y_i = (g + \delta_i)k_i \quad i = 1, \dots, n$$

and the steady-state *PPF* becomes

$$(9') \quad c = T[(g + \delta_1)k_1, \dots, (g + \delta_n)k_n; k_1, \dots, k_n]$$

We must introduce further properties of the *PPF*. First, it is assumed to be concave in the  $y_i$  and  $k_i$  and twice continuously differentiable; thus it has a negative semidefinite Hessian matrix,  $H = [T_{ij}]$ . Second, it is assumed to reflect an underlying technology in which there is no joint production. As Samuelson (1966a) has shown, this assumption imposes certain restrictions on the corresponding function  $T$ ; in particular, it implies that the rank of the Hessian matrix  $H$  cannot exceed  $n$ . However, since joint production raises further complications which we wish to avoid for the present, we shall delay a discussion of it until Section III. We introduce the following notation:

$p_0$  = equilibrium price of the consumption good in terms of the wage rate as numeraire,

$p_i$  = equilibrium price of the  $i$ th capital good in terms of the wage rate as numeraire,

$w_i$  = the gross rental rate for the  $i$ th capital good in terms of the wage rate as numeraire.

Then at every interior steady-state equilibrium, we have

$$(10) \quad \begin{cases} T_i = -p_i/p_0 \\ T_{n+i} = w_i/p_0 \end{cases} \quad i = 1, \dots, n$$

together with the conditions

$$(11) \quad r = w_i/p_i - \delta_i \quad i = 1, \dots, n$$

Conditions (10) are simply cost minimization requirements necessary for static ef-

iciency or profit maximization. Conditions (11) are necessary for competitive equilibrium; they insure that the net own-rate of return for every capital good, the right-hand side of (11), is equal to the exogenous rate of interest. Hence, it follows from (10) and (11) that

$$(12) \quad \frac{T_{n+i}}{T_i} = -(r + \delta_i) \quad i = 1, \dots, n$$

It has been shown elsewhere (see, for example, Burmeister and Kiyoshi Kuga, Burmeister and A. Rodney Dobell, ch. 9, and Michio Morishima), that under certain conditions the steady-state equilibrium quantities  $k_i$  are functions only of  $r$  for all values of  $r$  belonging to a non-empty open interval  $R$ . Likewise, for any assigned value of the steady-state interest rate  $r \in R$ , all steady-state equilibrium prices and the capital good rental rates are unique and positive, with the additional properties that  $dp_i/dr > 0$ ,  $dw_i/dr > 0$ ,  $i = 1, \dots, n$ .<sup>6</sup> Thus, all equilibrium quantities may be expressed as functions of  $r$  and, in particular, the steady-state *PPF* may be written as

$$(9'') \quad c(r) = T[(g + \delta_1)k_1(r), \dots, (g + \delta_n)k_n(r); k_1(r), \dots, k_n(r)]$$

for all  $r \in R$ .

To determine how steady-state consumption responds to a change in  $r$ , we differentiate (9'') with respect to  $r$ , yielding

$$(13) \quad \frac{dc(r)}{dr} = \sum_{i=1}^n [T_i(g + \delta_i) + T_{n+i}] \frac{dk_i(r)}{dr}$$

Substituting the equilibrium conditions (10) and (11), we can write (13) as

<sup>6</sup> We must impose some additional restrictions not mentioned here to assure that all steady-state prices and quantities are strictly positive and that the interval  $R$  is not empty. These technical details may be ignored for our present purposes, and the interested reader is referred to Burmeister and Dobell, Burmeister and Kuga, Morishima. As before, we assume  $g \in R$ .

$$(13') \quad \frac{dc}{dr} = (r - g) \sum_{i=1}^n (p_i/p_0)(dk_i/dr)$$

(This result is stated and proved as Theorem 7 in Burmeister and Dobell, p. 286.)

Since  $p_0 > 0$  for all  $r \in R$ , we conclude that  $\text{sgn}(dc/dr) = \text{sgn}(g - r)$ , excluding possible stationary points of the function  $c(r)$  which are not relative maxima or minima, *if and only if*

$$(14) \quad \sum_{i=1}^n p_i(dk_i/dr) \leq 0 \quad \text{for all } r \in R$$

Consequently, there exists no paradoxical consumption behavior (as we have defined it above) if, and only if, (14) holds, and we thus suggest the following definition.

#### DEFINITION

The neoclassical model described above is said to exhibit *capital deepening response* at a given interest or profit rate  $r$  if, and only if, the expression

$$\sum_{i=1}^n p_i(dk_i/dr),$$

which itself is a function of  $r$ , is non-positive for that value of  $r$ . The model is termed *regular* if this inequality holds for all values of  $r \in R$ .

(Note that when  $n=1$ , this definition is satisfied in models with one capital good, where  $dk_1/dr < 0$  always holds; see equation (6) above and the concluding sentence of Section I.) Thus, consumption behavior across steady-state equilibria is "well behaved"—i.e., there is no paradoxical consumption behavior—if, and only if, the economy is regular.

We now compare our proposed definition of capital deepening response with a more common definition. One natural extension of the property  $dk/dr < 0$  when there are many heterogeneous capital goods is the

requirement

$$(15) \quad \frac{dv}{dr} < 0$$

where

$$v \equiv \sum_{i=1}^n p_i k_i$$

and is the value of the per capita capital stock. This requirement is precisely the "unobtrusive postulate" which Pasinetti has attributed to neoclassical authors (1969, 1970) and which Robert Solow (1970) has denied.

Differentiating  $v$ , we observe that the change in the value of capital is

$$(16) \quad \begin{aligned} dv/dr &= \sum_{i=1}^n p_i(dk_i/dr) \\ &+ \sum_{i=1}^n (dp_i/dr)k_i \end{aligned}$$

Since the second term on the right-hand side of (16) is positive—it is a pure Wicksell effect—we know that

$$(17) \quad dv/dr < 0 \Rightarrow \sum_{i=1}^n p_i(dk_i/dr) < 0$$

Accordingly, when Pasinetti's "unobtrusive postulate" is satisfied, so is our definition of capital deepening response. Pasinetti's criticisms concerning this particular issue would be completely valid provided the implication could be reversed and (17) were an *if, and only if*, relationship. Unfortunately, a numerical example demonstrates conclusively that  $dv/dr$  may be *positive* even when

$$\sum_{i=1}^n p_i(dk_i/dr)$$

is always negative. (In such cases one can calculate an upper bound on  $dv/dr$  by differentiating the national income identity and using the steady-state restrictions.) Such an example is reported by Bur-

meister and Dobell, pp. 291-92.<sup>7</sup> The primary conclusion of our analysis is stated quite simply:

Although the *value of capital* may increase as the interest or profit rate increases, an economy may still be "well behaved" or regular. It is the *change in the per capita value of capital stocks evaluated at equilibrium prices* which must always remain nonpositive to preclude paradoxical consumption behavior.

The one-sector model *always* exhibits capital deepening response and is regular. In the multi-capital good case we have thus far established only the equivalence of our definition to that of no paradoxical consumption behavior; we have said nothing as to whether or not multi-capital good models are in fact regular. To answer this question we must derive explicit expressions for

$$\sum_{i=1}^n p_i (dk_i/dr)$$

Differentiating the equilibrium conditions (12) with respect to  $r$ , we obtain

$$(18) \quad \Omega \frac{dk}{dr} = -e$$

where

$$\frac{dk}{dr} = \left( \frac{dk_1}{dr}, \dots, \frac{dk_n}{dr} \right)',$$

$$e = (1, 1, \dots, 1)',$$

and now primes denote the transpose operator. The  $(n \times n)$  matrix  $\Omega$  is the Jacobian matrix of the system of equations (12) and is defined by  $\Omega = [\omega_{ij}]$  where the elements are given by (12.)\* We consider two cases: I.  $\Omega$  is nonsingular and has rank  $n$  for all

<sup>7</sup> Even in a well-behaved neoclassical two-sector model  $dv/dr = d(p_1 k_1)/dr$  may be positive provided the term  $(dp_1/dr)k_1$  is sufficiently positive to outweigh the negative term  $p_1(dk_1/dr)$ ; presumably, severe capital intensity conditions will yield this result.

$r \in R$ . II. The rank of  $\Omega$  is less than or equal to  $n$  for all  $r \in R$ . Clearly, Case II is more general and includes Case I.

Case I. Define the matrix

$$P = \text{diag}(T_1, \dots, T_n),$$

the equilibrium conditions (10) yield

$$P = -(1/p_0) \text{diag}(p_1, \dots, p_n)$$

Hence, we can use the optimality conditions (12) to find a relationship between  $\Omega$  and the Hessian matrix of  $T$ , denoted by  $H$ . Thus we calculate

$$(19) \quad \Omega = P^{-1}RHG'$$

where  $D$  is defined in (20) and  $G$  is defined analogously with  $g$  replacing  $r$  in (20)\*. Equations (18) and (19) together imply that<sup>8</sup>  $dk/dr = -[RHG']^{-1} \cdot Pe$ . But since  $Pe = -(1/p_0)p$ , where  $p = (p_1, \dots, p_n)'$ , we obtain the result

$$(21) \quad \sum_{i=1}^n p_i (dk_i/dr) \equiv p' \frac{dk}{dr} = (1/p_0) p' [RHG']^{-1} p$$

Thus, we see from (21) that our definition of capital deepening response will be satisfied in Case I if, and only if,

$$(22) \quad p' [RHG']^{-1} p \leq 0,$$

and (22) is satisfied if the matrix  $[RHG']$  (and hence  $[RHG']^{-1}$ ) is quasi-negative definite.<sup>9</sup> In general there is nothing to ensure that this condition will be met.

<sup>8</sup> Essentially in Case I, we assume  $\det [RHG'] \neq 0$  for all  $r \in R$ . This in turn requires that the rank of the Hessian matrix  $H$ , which under conditions of no-joint production cannot exceed  $n$ , must in fact equal  $n$ .

<sup>9</sup> The notion of negative definiteness applies only to symmetric matrices. The analogous notion for nonsymmetric matrices is that of quasi-negative definiteness. More precisely, a nonsymmetric matrix  $A$  is quasi-negative definite if, and only if, the symmetric matrix  $[A + A']/2$  is negative definite. Since  $x'A x = x'A' x$  for all  $x$ , we immediately see that quasi-negative definiteness is equivalent to  $x'A x < 0$  for  $x \neq 0$ .

\* See p. 848 for equations (12') and (20).

$$(12') \quad \omega_{ij} = \left[ \frac{T_i[(g + \delta_j)T_{n+i,j} + T_{n+i,n+j}] - T_{n+i}[(g + \delta_j)T_{ij} + T_{i,n+j}]}{T_i^2} \right]$$

$$(20) \quad R = \begin{bmatrix} (r + \delta_1) & 0 & \dots & 0 & 1 & 0 & 0 \\ 0 & (r + \delta_2) & \dots & 0 & 0 & 1 & 0 \\ \cdot & \cdot & & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \dots & (r + \delta_n) & 0 & 0 & \dots 1 \end{bmatrix}$$

However, it is worth mentioning that when  $r=g$ , the matrix  $R=G$  and the requirement (22) becomes

$$(22') \quad p'[GHG']^{-1}p \leq 0$$

Since it can be shown that  $c(r)$  attains a unique global maximum at  $r=g$ , equation (14), and hence (22) in Case I, will be satisfied with strict inequality in a deleted neighborhood of  $r=g$ ; thus it follows that the economy exhibits capital deepening response, at least locally in the neighborhood of the Golden Rule point.<sup>10</sup> This, of course, proves that near the Golden Rule

<sup>10</sup> Let  $x_i^*$  and  $x_i^{**}$  be two input vectors associated with the  $i$ th neoclassical production function  $F^i(x_i)$ ,  $i=0, 1, \dots, n$ , and suppose  $x_i^* \neq \mu x_i^{**}$  for any positive scalar  $\mu$ . Then by concavity

$$F^i[\lambda x_i^* + (1-\lambda)x_i^{**}] > \lambda F^i(x_i^*) + (1-\lambda)F^i(x_i^{**})$$

for all  $\lambda \in (0, 1)$ . Normalize on total labor by setting  $L=1$ , and denote the global maximum value of steady state per capita consumption by  $c^*$ . (Our assumptions ensure that such a  $c^*$  exists and that  $g \in R$ .)

Assume  $c^* = F^0(x_0^*)$  when  $r=g$  and suppose  $c^* = F^0(x_0^{**})$  for some other admissible  $\bar{r} \neq g$ . Clearly, it is impossible that  $x_0^* = \mu x_0^{**}$  for any positive scalar  $\mu$  because across steady states all relative factor prices must satisfy  $d(W_i/W_0)/dr = dw_i/dr > 0$ . Accordingly,  $c = c^*$  if, and only if,  $r=g$ , for otherwise we have an obvious contradiction to the strict concavity of  $F^0(x_0)$  for nonproportional input vectors  $x_0$ . Thus, there exists a neighborhood of the point  $r=g$ , say  $N(g)$ , such that  $d^2c(r)/dr^2 \leq 0$  for all  $r \in N(g)$  with strict inequality if  $r \neq g$ . The statement in the text then follows immediately from 1) the existence and continuity of the

equilibrium a multi-capital good model is "well-behaved," as would be expected.

By rewriting equation (13') in the form

$$(13'') \quad p_0(r) \frac{dc(r)}{dr} + g \sum_{i=1}^n p_i(r) \frac{dk_i(r)}{dr} = r \sum_{i=1}^n p_i(r) \frac{dk_i(r)}{dr}$$

we see how our results relate to some recent work of von Weizsäcker, pp. 62-63. Recalling that the steady-state output of capital goods is given by equation (3''), we can interpret the left-hand side of (13'') as being the marginal change in the value of net national product, while the right-hand side is equal to  $r$  times the marginal changes in (per capita) capital stocks weighted by equilibrium prices. Thus we confirm von Weizsäcker's result: "... if measured in constant prices, the marginal productivity of capital is equal to the rate of interest" (p. 62). This result is the direct analogue of equation (5) obtained for the one-sector model and thus provides an additional argument in favor of our definition of capital deepening re-

steady-state functions  $k_i = k_i(r)$ ,  $r \in R$ ,  $i=1, \dots, n$ , and 2) the equation

$$d^2c(r)/dr^2 = \sum_{i=1}^n (p_i/p_0)(dk_i/dr) \quad \text{at } r = g$$

sponse. Furthermore, at  $r=g$

$$(23) \quad p_0(r) \frac{d^2 c(r)}{dr^2} + g \sum_{i=1}^n p_i(r) \frac{d^2 k_i(r)}{dr^2} \leq r \sum_{i=1}^n p_i(r) \frac{d^2 k_i(r)}{dr^2}$$

which von Weizsäcker interprets as a *local* law of diminishing marginal returns to capital. This result seems logically equivalent to our statement that a multi-capital good model is well-behaved near  $r=g$ ; we relegate the details of this observation to a footnote.<sup>11</sup>

We now turn our attention to the general case in which the rank of  $\Omega$  is less than or equal to  $n$ .

Case II. In this case the matrix  $RHG'$  is in general singular and we are unable to obtain equation (21) and inequality (22). We thus must adopt a somewhat different strategy, but one which provides additional insights. From our assumption of no-joint production, each industry possesses its own individual production function all of which underlie the aggregate *PPF*. By examining these individual production functions, we can show that, despite the fact (14) may not be satisfied, it nevertheless seems likely that neo-classical multi-capital good model will be regular except in unusual circumstances.

Gross outputs are determined by the neoclassical production functions

$$(24) \quad Y_j = F^j(L_j, K_{1j}, \dots, K_{nj}) \quad j = 0, 1, \dots, n$$

<sup>11</sup> Define the row vector  $p \equiv (p_1, \dots, p_n)$  and the column vector  $k \equiv (k_1, \dots, k_n)$ . Denoting derivatives with respect to  $r$  by primes, we may write (13') as  $p_0 c' = (r-g)p k'$ . Differentiating the latter again with respect to  $r$  yields  $p_0 c'' + p_0 c' = (r-g)(p k'' + p' k') + p k'$  which reduces to  $p_0 c'' + g p k' = r p k'' + \alpha$ , where  $\alpha \equiv (r-g)p' k' + p k' - p_0 c'$ . Equation (23) is verified provided  $\alpha \leq 0$  at  $r=g$ , which is equivalent to  $p k' \leq 0$  at  $r=g$ . However, we have already proved that  $p k' \leq 0$  at  $r=g$ ; see fn. 10.

where  $L_j$  and  $K_{ij}$  are the inputs of labor and capital of type  $i$  employed producing good  $j$ .<sup>12</sup> For uniformity of notation, the consumption good industry is now indexed by  $j=0$ . We also introduce the notation

$$k_{ij} = K_{ij}/L_j \quad \text{and} \\ \rho_j = L_j/L$$

hence,  $k_{ij}$  denotes the capital-labor ratio of capital good  $i$  in the  $j$ th industry and  $\rho_j$  denotes the proportion of the total labor force employed in that industry. We define the aggregate quantities

$$K_i = \sum_{j=0}^n K_{ij} \quad \text{and} \\ L = \sum_{j=0}^n L_j$$

and thus we find

$$(25) \quad k_i = \sum_{j=0}^n k_{ij} \rho_j \\ 1 = \sum_{j=0}^n \rho_j$$

We also denote the aggregate measure of capital deepening response by

<sup>12</sup> Analogous to fn. 3 above,  $F^j_i$  denotes the partial derivative of  $F^j$  with respect to its  $i$ th argument and is assumed to be positive while the Hessian matrix  $[F^j_{ik}]$  is negative semi-definite. Also,  $F^j$  is assumed homogeneous of degree one for all  $j$  and twice continuously differentiable. We also suppose that  $L_j, K_{ij} > 0$  for all  $i, j$ , although this assumption is merely a simplifying convenience and can be weakened in many ways. If the underlying technology consists of many fixed-coefficient Leontief techniques, it is still true that the economy has a *PPF* of the form (9), although, of course, it is not differentiable. However, as Samuelson has pointed out to us, at switching points between alternative techniques there exist supporting hyperplanes with slopes exactly analogous to the partial derivatives of  $T$ ; see Samuelson (1966a, p. 40) for a definition of such "generalized partial derivatives." Likewise, with linear technologies the individual production functions are not differentiable, but again there exist supporting hyperplanes with slopes corresponding to "generalized partial derivatives." These facts allow a straightforward extension of the analysis given here to cover the linear case.

$$\theta \equiv \sum_{i=1}^n p_i (dk_i/dr)$$

and analogously define

$$\theta_j \equiv \sum_{i=1}^n p_i (dk_{ij}/dr) \quad j = 0, 1, \dots, n$$

as the measure of capital deepening response in the  $j$ th industry (expressed in terms of *its own* capital-labor ratios). We shall now express  $\theta$  in terms of the individual sectoral components.

Because of the homogeneity of  $F^j$  we can write the production functions (24) in the intensive form

$$(24') \quad Y_j/L_j = f^j(k_{1j}, \dots, k_{nj}) \quad j = 0, 1, \dots, n$$

where by assumption  $f^j$  is strictly concave and

$$H^j = [f_{ik}^j]$$

is negative definite. The equilibrium conditions for industry  $j$  require that

$$(26) \quad f_{ik}^j(k_{1j}, \dots, k_{nj}) = (r + \delta_i) p_i / p_j \quad i = 1, \dots, n$$

Note that equations (26) are a generalization of equation (5).

To see how the  $k_{ij}$  respond to change in  $r$ , we differentiate the set of equations (26) with respect to  $r$ , yielding

$$(27) \quad [H^j] \frac{dk_{.j}}{dr} = \frac{1}{p_j} p + (r + \delta) \phi_j \quad j = 0, 1, \dots, n$$

where

$$\begin{aligned} \frac{dk_{.j}}{dr} &= \left( \frac{dk_{1j}}{dr}, \dots, \frac{dk_{nj}}{dr} \right)' \\ r + \delta &= \text{diag}(r + \delta_1, \dots, r + \delta_n) \\ \phi_j &= \left( \frac{d}{dr} (p_1/p_j), \dots, \frac{d}{dr} (p_n/p_j) \right)' \end{aligned}$$

Hence because of the nonsingularity of  $H^j$  it follows that

$$(28) \quad \theta_j \equiv p' \frac{dk_{.j}}{dr} = \frac{1}{p_j} p' [H^j]^{-1} p + p' [H^j]^{-1} (r + \delta) \phi_j$$

The negative definiteness of  $H^j$  implies that the first term on the right-hand side of (28) is negative. However, since the relative prices of capital goods in terms of  $p_j$  can either increase or decrease in response to a change in  $r$ , the second term is ambiguous, making the overall sign of  $\theta_j$  indeterminate. In the special case where the responses of relative prices to changes in the interest rate are small, we can expect the first term to dominate, implying that  $\theta_j < 0$ . Hence, in that case we reach the conclusion that *each sector* will exhibit capital deepening response and will be regular (as we have defined these terms) *if intensities are expressed in terms of the sectoral capital-labor ratios*. However, even this special case does not ensure capital deepening response or regularity in the aggregate.

We can calculate the aggregate measure of capital deepening response,  $\theta$ , by differentiating (25) with respect to  $r$ :

$$(29) \quad \begin{aligned} dk_i/dr &= \sum_{j=0}^n \rho_j (dk_{ij}/dr) \\ &+ \sum_{j=0}^n k_{ij} (d\rho_j/dr), \end{aligned}$$

where

$$\sum_{j=0}^n (d\rho_j/dr) = 0$$

Hence, we derive the expression

$$\begin{aligned} \theta &\equiv \sum_{i=1}^n p_i (dk_i/dr) = \\ &\sum_{i=1}^n p_i \left[ \sum_{j=0}^n \rho_j (dk_{ij}/dr) + \sum_{j=0}^n k_{ij} (d\rho_j/dr) \right] \end{aligned}$$

so that

$$(30) \quad \theta = \sum_{j=0}^n \rho_j \theta_j + \sum_{j=0}^n v_j (d\rho_j/dr)$$

where

$$v_j \equiv \sum_{i=1}^n p_i k_{ij} > 0$$

denotes the value of the capital stock in industry  $j$  per worker employed in that sector. Thus in the case that  $\theta_j < 0$ , the first summation on the right-hand side of (30) is always negative. However, in view of (29), the sign of the second term of (30) is ambiguous, thereby raising the possibility that even in this case the model may not be regular in the aggregate. We may note there are two cases in which the second term vanishes: If all the  $v_j$ 's are the same; If  $d\rho_j/dr = 0$  for all  $j$ , in which case there are no absolute movements of labor in response to changes in  $r$ .

In general, since the summation

$$\sum_{j=0}^n v_j (d\rho_j/dr)$$

necessarily includes terms of opposite signs, it seems likely that they will tend to cancel out, in which case the first term of (30) will dominate. Also if the capital deepening responses within each sector are strong, it becomes more likely that aggregate deepening response will prevail; that is, the larger (in absolute value) the negative magnitude of each  $\theta_j$ , the more likely it is that the first term of (30) will dominate and  $\theta$  will be negative. On the other hand, even if some  $\theta_j$ 's are positive  $\theta$  may still be negative, provided other  $\theta_j$ 's are sufficiently negative. And  $\theta$  is almost certainly negative if the  $v_j$ 's are approximately the same in each sector and each  $\theta_j < 0$ . In other words we see that if each

sector is regular yet the model is not regular in the aggregate so that paradoxical consumption behavior exists, it would require an industry, say industry  $j$ , in which  $v_j$  is large relative to that of the other industries and where  $d\rho_j/dr > 0$ . Under these circumstances an increase in  $r$  would cause a shift in the proportion of labor towards an industry with a high per capita value of capital. This shift would increase the aggregate per capita value of capital in the economy and possibly (but not necessarily) would lead to a violation of our definition of a regular model; see equation (16).

In summary, we see that within each sector our definition of capital deepening response may or may not be satisfied. However, even if each sector is regular, this does not ensure the model will be regular in the aggregate; that will depend upon the relative capital intensities of the different industries and how labor shifts between them in response to interest rate changes. If the sectoral capital deepening response effects are sufficiently strong, however, the economy is likely to be regular in the aggregate, and in any event, it will always show local capital deepening response in the neighborhood of the Golden Rule point  $r=g$ . Finally, our results suggest an interesting conjecture far beyond the scope of this paper. If one argues, as many economists do, that industrialized economies moving over time are always "near" steady-state equilibria positions, then one possible reason that an "aggregate production function" gives a good fit to the data may be because econometricians have studied economics with underlying technologies that are regular (in our terminology).

### III. The Case of Joint Production

Under conditions of joint production the society's PPF will still be of the form (9) although, as we have shown elsewhere (see

Burmeister and Turnovsky), the function  $T$  will be subjected to certain restrictions analogous to those obtained by Samuelson (1966a) for the no-joint production case. The steady-state equilibrium conditions are still given by (10) and (11) although now a severe difficulty arises. Unlike the no-joint production case, it is no longer true that the equilibrium steady-state values of the  $k_i$ 's are determined only by  $r$ . Thus, the  $PPF$  cannot necessarily be written in the form (9''). Instead we now face the possibility that for each value of  $r$  (at least over some range) we may be able to associate a *number* of different  $k_i$ . Specifically, suppose that when  $r=r_1$ , there exist three vectors  $k^i=(k_1^i, \dots, k_n^i)$ ,  $i=1, 2, 3$ , which are consistent with the equilibrium conditions (12); then there will be three corresponding values of consumption  $c^i$  which satisfy (9). In this case we may apply our definition of capital deepening response to each of the three equilibrium values separately. Hence, we say that the model exhibits capital deepening response in the neighborhood of the equilibrium pair  $(r_1, k^i)$ , if, and only if,

$$(31) \quad \sum_{i=1}^n p_i \left. \frac{dk_i}{dr} \right|_{r=r_1, k=k^i} \leq 0$$

Furthermore, it readily follows by our previous argument that

$$\left. \text{sgn} \frac{dc^i}{dr} \right|_{r=r_1, c=c^i, k=k^i} = \text{sgn}(g-r)$$

and there is no paradoxical consumption behavior in the neighborhood of this particular equilibrium *if, and only if*, (31) holds.

Since the difficulties introduced by joint production arise even with a single capital good, it is easy to consider that special case in more detail. The steady-state  $PPF$  becomes

$$(32) \quad c = T[(g+\delta)k; k]$$

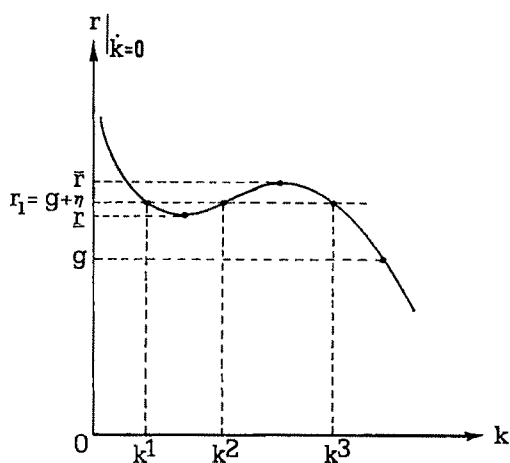


FIGURE 1

while the steady-state equilibrium pricing conditions imply

$$(33) \quad r = -T_2/T_1 - \delta$$

Differentiating both sides of (33) with respect to  $k$ , we find

$$(34) \quad \text{sgn} \frac{dr}{dk} = \text{sgn} \{ -T_1[T_{21}(g+\delta) + T_{22}] + T_2[T_{11}(g+\delta) + T_{12}] \}$$

In our model  $r$  is exogenous, and when  $r=g$ , we obtain  $T_2/T_1 = -(g+\delta)$ . Substituting into (34) and using the strict concavity of  $T$ , we deduce that

$$(35) \quad \left. \frac{dk}{dr} \right|_{r=g, k=0} < 0$$

By a similar substitution, it can be readily verified that

$$\left. \frac{dk}{dr} \right|_{r_1=g+\eta, k=0} > 0$$

is a possibility if  $(r_1-g)$ , which we denote by  $\eta$ , is sufficiently large and  $T_{12} > 0$ .<sup>13</sup>

<sup>13</sup> See Liviatan and Samuelson. Apparently this possibility was first observed by Richard Sutherland in an unpublished Ph.D. dissertation. Subsequent work by

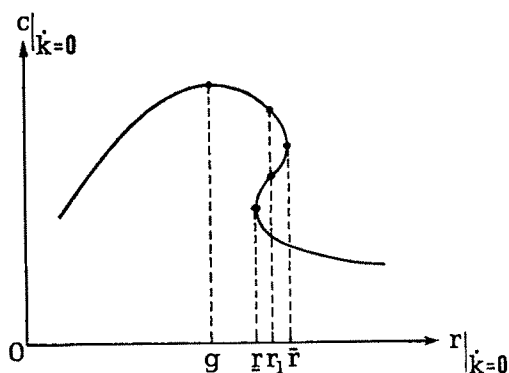


FIGURE 2

Plotting  $r$  against  $k$ , we obtain Figure 1 where "waves" occur in the neighborhood of  $r_1 = (g + \eta)$ .

Now since steady-state consumption is given by (32), we can use  $k$  as a parameter to plot  $c$  versus  $r$  in Figure 2. Thus, even in neoclassical models with only one capital good, we may have paradoxical consumption behavior for profit or interest rates larger than  $g$ , provided we admit the possibility of joint production. Note, however, that if our definition of capital deepening response is satisfied at *every* (admissible) steady-state interest rate  $r$ , then the existence of "multiple consumption turnpikes" in joint production models is precluded. This observation is significant because only models with a unique consumption turnpike exhibit qualitative turnpike properties which are independent of the initial capital stocks.<sup>14</sup> Accordingly, the concept of a regular economy which we have introduced in this paper apparently has applications to economic problems not considered here.

others has dealt with related issues which are beyond the scope of this paper.

<sup>14</sup> See Burmeister and Dobell, pp. 411-16, for an elementary discussion of this problem.

## REFERENCES

- M. Bruno, E. Burmeister, and E. Sheshinski, "The Nature and Implications of the Re-switching of Techniques," *Quart. J. Econ.*, Nov. 1966, 80, 526-53.
- E. Burmeister, "The Existence of Golden Ages and Stability in the Two-Sector Model," *Quart. J. Econ.*, Feb. 1967, 81, 146-54.
- and S. J. Turnovsky, "The Degree of Joint Production," *Int. Econ. Rev.*, Feb. 1971, 12, 99-105.
- and A. R. Dobell, *Mathematical Theories of Economic Growth*, New York 1970.
- and K. Kuga, "The Factor-Price Frontier in a Multi-Sector Neoclassical Model," *Int. Econ. Rev.*, Feb. 1970, 11, 162-74.
- N. Liviatan, and P. A. Samuelson, "Notes on Turnpikes: Stable and Unstable," *J. Econ. Theor.*, Dec. 1969, 1, 454-75.
- M. Morishima, *Equilibrium, Stability, and Growth*, Oxford 1964.
- L. Pasinetti, "Switches of Technique and the 'Rate of Return' in Capital Theory," *Econ. J.*, Sept. 1969, 79, 508-31.
- , "Again on Capital Theory and Solow's Rate of Return," *Econ. J.*, June 1970, 80, 428-31.
- J. Robinson, *The Accumulation of Capital*, Homewood 1956.
- P. A. Samuelson, (1966a) "The Fundamental Singularity Theorem for Non-Joint Production," *Int. Econ. Rev.*, Jan. 1966, 7, 34-41.
- , (1966b) "A Summing Up," *Quart. J. Econ.*, Nov. 1966, 80, 568-83.
- R. M. Solow, "The Interest Rate and Transition Between Techniques," in C. H. Feinstein, ed., *Socialism, Capitalism, and Economic Growth*, New York 1967, 30-39.
- , "On the Rate of Return: Reply to Pasinetti," *Econ. J.*, June 1970, 80, 423-28.
- R. Sutherland, "On Optimal Development Programs when Future Utility is Discounted," unpublished doctoral dissertation, Brown Univ. 1967.
- C. C. von Weizsäcker, *Steady State Capital Theory*, New York 1971.

# Interest Rates and Inflationary Expectations: New Evidence

By WILLIAM E. GIBSON\*

The effects of expected inflation on market interest rates have concerned economists for decades, particularly in recent years when the relationship has been judged to have especially important empirical relevance. Irving Fisher's original description of the problem has been generally convincing on the theoretical level, although some controversy remains regarding the effects of price expectations on the real rate of interest (see Reuben Kessel and Armen Alchian, Robert Mundell, Michael Porter). Most of the recent work on the subject has been empirical (see William Gibson, Thomas Sargent, William Yohe and Denis Karnosky). Several techniques have been involved, but all attempt to measure the relationship indirectly via hypotheses regarding the generation of expectations. The most common assumption used has been that price expectations are based on past price experience. Such constructions have been made necessary by the absence of directly observed data on price expectations in the United States. This paper seeks to overcome these difficulties by employing previously unused (for this purpose) data on expectations obtained by direct surveys of important market participants.

## I. The Data

Joseph Livingston, a nationally syndicated financial columnist, has twice yearly

\* Senior staff economist, Council of Economic Advisers. A portion of the work reported here was undertaken while I was at the Federal Deposit Insurance Corporation. Myron Kwast and H. Kemble Stokes provided valuable research assistance for the preparation of this paper.

since 1946 surveyed a group of business, government, labor, and academic economists on their expectations of future values of selected aggregate economic variables. The latter include the consumer price (cost of living) index, six and twelve months hence. The early respondents were primarily business and financial economists; the coverage has been broadened through the years to include also well-known government and academic economists. Livingston has selected the sample himself and has sought individuals who "are in strategic positions to influence decisions of businessmen." A complete listing of the respondents in 1947 and 1970 appears as an Appendix to this paper.

From these data, expected rates of change of prices can be constructed, and the statistical construct can be related to market interest rates. It should be noted that this approach can only approximate the true relationship, which depends on the sum of the actions of all market participants, not merely those of a select few. Thus, the expectations surveyed include only a portion of the relevant variable. But these expectations may well approximate those of the market as a whole, since the sample was chosen to include well-informed individuals. The group selected also includes those whose opinions carry much greater weight in forming market expectations than those of an ordinary citizen, for the members of the sample typically control large amounts of assets and can thus have fairly large influence on the markets in which expectations have effects.

## II. Expectations and Interest Rates

When inflation becomes expected, lenders expect the real value of their principal and interest payments to be depreciated and borrowers expect to be able to repay loans with money for which less real value must be sacrificed than before expectations changed. Thus at any level of market interest rates the quantity of loans supplied decreases while the quantity demanded increases. Both forces increase nominal interest rates. If the real rate of interest (the nominal rate less compensation for expected inflation, or  $r = i - \dot{p}^*$ ) remains unchanged, the nominal rate will rise by the increase in the expected rate of inflation. If the real rate falls when the expected rate of inflation increases, the nominal rate will rise by less than the expected rate by the extent of the decline in the real rate. There is as yet no theoretical consensus on the relationship between the real rate and the expected rate of inflation.<sup>1</sup> As a result, it is assumed below that variations in price expectations do not affect the real rate. The empirical measures developed here do, however, provide some evidence on this question.

We can measure the effects of price expectations on interest rates in the United States by estimating

$$(1) \quad i = a_0 + a_1 \dot{p}^*,$$

where  $i$  is the nominal rate of interest and  $\dot{p}^*$  is the expected rate of price change. Equation (1) does not represent a full and complete theory of interest rate determination. Rather it seeks merely to quantify the effect of expectations on nominal interest rates.

Several *a priori* assertions can be made about the values  $a_1$  should take in equation (1). First, there is reason to believe that  $a_1$  should be between zero and one. A value of

$a_1 = 1$  is consistent with a world in which the real rate of interest is unaffected by price expectations and nominal interest rates fully adjust to these expectations. This result is, however, also consistent with a world in which positive (negative) effects on the real rate are exactly matched by underadjustment (overadjustment) of nominal rates. No theoretical arguments have been advanced that these latter outcomes should always hold, but they remain possibilities. Similarly,  $a_1 = 0$  is consistent with an unchanged real rate and no adjustment of the nominal rate to expectations as well as with other combinations of real and nominal rate adjustments.

Second, the measures used here should affect interest rates differently depending on the term to maturity of the security. For instance, interest rates on loans for six months should respond to expectations of price movements over the coming six months but not to expectations for periods after that, for price movements after six months have no effect on the outcome in real terms for the borrower or lender. Similarly, six-month expectations should also influence yields on loans of longer than six months, for the coming six months is included in the longer period. In addition, market participants may have greater certainty of the accuracy their first six-month expectations than of those for more distant periods. Since the first six months is a smaller portion of the term of a loan the longer the loan, we would also expect the influence of six-month expectations to diminish as term to maturity increases beyond six months. The same considerations hold for the effects of one-year expectations on rates for longer than a year.

## III. Estimated Effects

In order to eliminate effects of changing default risk premiums, yields on U.S. Treasury securities were used to measure

<sup>1</sup> See Kessel and Alchian, Mundell, and Porter.

market interest rates.<sup>2</sup> Five different maturity categories were used, ranging from 3-month bills to 10-year and longer term to maturity bonds. All were obtained from the *Federal Reserve Bulletin*. Market yields were used rather than new issue yields, and the post-Accord period was selected to cover a period of interest rate flexibility. Estimates of equation (1) for various yields, the 6-month and 12-month expected rates of inflation appear in Table 1. Every coefficient in this table is estimated to be significantly positive at the 0.05 level or better, indicating a strong association between interest rates and this measure of expectations. Interest rates are shown to respond quickly to changes in expectations: although the response may take up to six months, this is far less than the lags usually found using weighted averages of past price changes to approximate expectations. We would expect a coefficient of 1.0 if there were perfect adjustment and no change in the real rate, i.e., a one percentage point increase in the expected rate of inflation leads to a one percentage point increase in the rate of interest. Over half the coefficients come quite close to this result.<sup>3</sup> The coefficients also vary with term to maturity as hypothesized. The 6- and 12-month expected rates of inflation have their largest effects on the 6-month and 9-12 month bill rates, and the coefficients decline as term to maturity increases. Although the long-term rate should be influenced by expectations covering a very long period, it is increased by 0.450 of the change in the rate expected for the following 6 months and by 0.675 of the rate ex-

TABLE 1—INTEREST RATES AND EXPECTED  
RATES OF INFLATION  
1952-1970  
(Observations at 6-month intervals;  
end of June and December)

Rate	Constant	$\hat{p}_{t+n}^*$	$R^2$	S.E.
$n=6$				
3-month bills	2.643 <sup>a</sup> (0.177)	0.6616 <sup>a</sup> (0.0764)	0.667	0.9535
6-month bills <sup>b</sup>	2.359 <sup>a</sup> (0.293)	0.9358 <sup>a</sup> (0.1140)	0.751	0.7578
9-12-month bills <sup>b</sup>	2.496 <sup>a</sup> (0.273)	0.9110 <sup>a</sup> (0.1062)	0.767	0.7061
3-5-year notes	3.371 <sup>a</sup> (0.164)	0.6113 <sup>a</sup> (0.0707)	0.666	0.8822
10-year and longer bonds	3.580 <sup>a</sup> (0.125)	0.4503 <sup>a</sup> (0.0540)	0.649	0.6742
$n=12$				
3-month bills	2.207 <sup>a</sup> (0.170)	0.9300 <sup>a</sup> (0.0854)	0.761	0.8076
6-month bills <sup>b</sup>	2.045 <sup>a</sup> (0.304)	1.0958 <sup>a</sup> (0.1236)	0.779	0.7174
9-12-month bills <sup>b</sup>	2.192 <sup>a</sup> (0.282)	1.0658 <sup>a</sup> (0.1149)	0.794	0.6637
3-5-year notes	2.921 <sup>a</sup> (0.133)	0.8959 (0.0668)	0.829	0.6317
10-year and longer bonds	3.230 <sup>a</sup> (0.094)	0.6750 <sup>a</sup> (0.0472)	0.847	0.4460

Source: Expected prices from Livingston; observed consumer price index and interest rates from *Federal Reserve Bulletin*.

<sup>a</sup> Denotes significant at the 0.05 level.

<sup>b</sup> The 6-month and 9- to 12-month bill rate series begin in 1959; sample here begins with 1959:12.

<sup>2</sup> Most of the estimations were also conducted for a variety of other widely quoted market rates, such as the commercial paper rate and various corporate bond yields, and the results were broadly consistent with those reported here.

<sup>3</sup> As will be seen more clearly below, the coefficients of the 6-month and 9- to 12-month bill rates are higher than the others partly because they are observed for a shorter period. Both of the series used here begin in 1959.

pected for the following 12 months. This result is consistent with other findings (Gibson, Sargent, Yohe and Karnosky) which suggest that long-term expectations are based heavily (but not to the same extent) on the same factors determining shorter term expectations. This is plausible, for shorter term price behavior can likely be predicted with greater accuracy than more distant experience, and the shorter term forms a part of the longer term.

Coefficients for the 12-month expected rate display roughly the same pattern as those of the 6-month rate, and exceed the latter for interest rates of one year and longer. This conforms with our anticipations since 12 months constitute a larger portion of the total periods for which expectations must be formulated. It is somewhat surprising, however, that the estimated coefficients of the 12-month expected rate of inflation on the 3- and 6-month bill rates are higher than those of the 6-month expected rate. The 12-month expected rate contains more information which is irrelevant to the terms to maturity of these rates than does the 6-month expected rate. The difference is not awesomely large for the 6-month bill rate, but for the 3-month bill rate the coefficient for the 12-month expectations is 41 percent higher than that for 6-month expectations. This result is difficult to explain, but it may be that holders of 3-month bills actually have a holding horizon well beyond three months. That is, they may when buying bills plan at the same time to roll the bills over several times even if other rates vary. If so, one-year expectations would be more relevant for their behavior. If holders of 6-month bills bought these instead for the convenience of the specific maturity in their portfolios, the ordering of the coefficients observed here would result. The very high level of the  $R^2$ s is somewhat surprising in light of all the other influences commonly listed as determining interest rate movements (for example, liquidity effects, income effects, congestion of the new issues calendar, tax dates, etc.). It appears that price expectations have a very strong influence on interest rates, even over a period less than a year.<sup>4</sup>

#### IV. Changes in Expectations Effects

Although the actual and expected rates of inflation have varied substantially since 1952, there is no particular reason to assume that the expectations-generating function or the effects of expectations on interest rates changed during the period. When the expected rate of inflation is low, we should find correspondingly small effects on interest rates, but not in a proportion different from those of larger expected rates. While no evidence on the stability of the relationship over time was sought, the unavailability of two of the interest rate series before 1959 inadvertently provided information on this point. Since coefficients in Table 1 tend to be higher for 6-month and 9- to 12-month maturities, the hypothesis was suggested that the difference stemmed from the difference in observation periods. To check this, the estimates of Table 1 were repeated for the period 1959:12 through 1970:12, the longest period for which all five yields are available. These results appear in Table 2 and show that the relationship is indeed sensitive to the estimation period. The estimates of  $a_1$  generally increase in the later period, as do the  $R^2$ s. The increases are particularly noticeable for the effects of six-month expectations: the coefficient for the shortest rate rose from 0.6616 to 0.9266, while that for the longest rate rose from 0.4503 to 0.6012. The higher coefficients also suggest that the adjustment of interest rates to expectations has been more complete during the later period than estimated for the period as a whole. Indeed, in this later period, adjustment has roughly been full, if we assume that expectations have not affected the real rate. The coefficients are very close to 1.0, im-

<sup>4</sup> It should be noted that these estimates of inflation were made concurrently with the interest rates to which they are here related. This fact should not lead to spurious correlation unless the respondents based their estimates of the price level directly on current changes in

interest rates. However, Livingston's columns amply document the many other unrelated factors determining these estimates. Other evidence (Stephen Turnovsky) suggests that the estimates are also related to past price experience.

TABLE 2—INTEREST RATES AND EXPECTED  
RATES OF INFLATION  
1959–1970(Observations at 6-month intervals; end of  
June and December, 1959:12–1970:12)

Rate	Constant	$\hat{p}_{t+n}^*$	$R^2$	S.E.
$n=6$				
3-month bills	2.204 <sup>a</sup> (0.302)	0.9266 <sup>a</sup> (0.1172)	0.737	0.7790
6-month bills	2.359 <sup>a</sup> (0.293)	0.9358 <sup>a</sup> (0.1140)	0.751	0.7578
9–12-month bills	2.496 <sup>a</sup> (0.273)	0.9110 <sup>a</sup> (0.1062)	0.767	0.7061
3–5-year securities	3.087 <sup>a</sup> (0.238)	0.8312 <sup>a</sup> (0.0924)	0.784	0.6140
10-year and longer bonds	3.437 <sup>a</sup> (0.177)	0.6012 <sup>a</sup> (0.0689)	0.774	0.4579
$n=12$				
3-month bills	1.889 <sup>a</sup> (0.312)	1.0869 <sup>a</sup> (0.1269)	0.767	0.7330
6-month bills	2.045 <sup>a</sup> (0.304)	1.0957 <sup>a</sup> (0.1237)	0.779	0.7141
9–12-month bills	2.192 <sup>a</sup> (0.282)	1.0658 <sup>a</sup> (0.1149)	0.794	0.6637
3–5-year securities	2.771 <sup>a</sup> (0.222)	0.9903 <sup>a</sup> (0.0905)	0.844	0.5226
10-year and longer bonds	3.170 <sup>a</sup> (0.144)	0.7342 <sup>a</sup> (0.0847)	0.877	0.3377

Source: Expected prices from Livingston; observed consumer price index and interest rates from *Federal Reserve Bulletin*.

<sup>a</sup> Denotes significant at 0.05 level.

plying that a one percentage point increase in the expected rate of inflation raises nominal interest rates by one percentage point.

The original estimates were also repeated for the period 1952:6 through 1959:6 for the three rates observed over this period. These results appear in Table 3 and confirm the difference in the relationship described above. The estimated  $a$ 's and  $R^2$ 's are uniformly lower, and the ordering of the coefficients conforms to our expectations.

TABLE 3—INTEREST RATES AND EXPECTED  
RATES OF INFLATION  
1952–1959(Observations at 6-month intervals; end of  
June and December, 1952:6–1959:6)

Rate	Constant	$\hat{p}_{t+n}^*$	$R^2$	S.E.
$n=6$				
3-month bills	2.254 <sup>a</sup> (0.206)	0.2587 <sup>a</sup> (0.1101)	0.244	0.7736
3–5-year securities	2.882 <sup>a</sup> (0.189)	0.1796 (0.1012)	0.133	0.7115
10-year and longer bonds	3.137 <sup>a</sup> (0.117)	0.0910 (0.0635)	0.070	0.4467
$n=12$				
3-month bills	2.158 <sup>a</sup> (0.208)	0.4537 <sup>a</sup> (0.2249)	0.180	0.8060
3–5 year securities	2.821 <sup>a</sup> (0.170)	0.4479 <sup>a</sup> (0.1831)	0.263	0.6562
10-year and longer bonds	3.108 <sup>a</sup> (0.105)	0.2578 <sup>a</sup> (0.1135)	0.229	0.4067

Source: Expected prices from Livingston; observed consumer price index and interest rates from *Federal Reserve Bulletin*.

<sup>a</sup> Denotes significant at 0.05 level.

Since this difference in coefficients by periods was discovered by chance, there is a good possibility that the shift in the relationship took place sometime other than 1959 and only happened to appear when the period was segmented in this particular way by data unavailability. Specifically, if a shift were to occur, one might expect it to appear around 1965 as inflation became more pronounced. This sharp increase and sustained nature of inflation might somehow have changed the public's responsiveness to inflation and expectations. To determine when the shift actually took place, the estimations in Tables 2 and 3 were repeated for various other points of time period separation between 1957:6 and 1966:6. The estimates of  $a_1$  for the 6-month and 12-month expected inflation rate over these subperiods for the 3-month Treasury bill rate are presented in Tables 4 and 5, in which the

TABLE 4—INTEREST RATES AND 6-MONTH EXPECTED RATES OF INFLATION FOR VARIOUS SUBPERIODS, 1952–1970

$$i = a_0 + a_1 \dot{p}_{t+6}^*$$

$$i = 3\text{-Month Treasury Bill Rate}$$

t	Subperiod 1952:6 through t			Subperiod t+6 through 1970:12		
	$a_1$	Standard Error of $a_1$	$R^2$	$a_1$	Standard Error of $a_1$	$R^2$
1957:6	.2462	.09697	.353	.8933	.10518	.732
1957:12	.2566	.09915	.341	.9185	.10864	.738
1958:6	.2466	.11050	.249	.8672	.10975	.719
1958:12	.2575	.10535	.277	.8599	.11357	.710
1959:6	.2587	.11005	.244	.9266	.11720	.737
1959:12	.3270	.11789	.309	.9330	.11819	.745
1960:6	.3217	.11259	.309	.9148	.12144	.736
1960:12	.3155	.10874	.304	.8884	.12544	.721
1961:6	.3072	.10461	.298	.8561	.12470	.719
1961:12	.2999	.09926	.300	.8320	.11675	.745
1962:6	.3012	.09564	.308	.8021	.11795	.739
1962:12	.3022	.09132	.321	.7739	.10979	.764
1963:6	.3055	.08997	.324	.7537	.12010	.733
1963:12	.3139	.09100	.322	.7558	.13369	.704
1964:6	.3242	.09037	.331	.7347	.14700	.667
1964:12	.3381	.09172	.335	.7369	.16712	.626
1965:6	.3404	.09455	.315	.8613	.21538	.600
1965:12	.3613	.09793	.318	1.0756	.26056	.640
1966:6	.3923	.09402	.370	1.0492	.28666	.608

Source: Expected prices from Livingston; observed consumer price index and interest rates from *Federal Reserve Bulletin*.

relevant estimates from Tables 2 and 3 are repeated for comparison purposes. The coefficients for other available interest rates show similar patterns. Maximum responsiveness of interest rates to expectations should be reflected in highest estimates of  $a_1$ .

When the post-1964 period is analyzed separately, the coefficients rise sharply in 1966: the jump in the estimate when the first point in the estimation period is advanced from 1956:6 to 1965:12 is in each case the largest on the table. The estimates of  $a_1$  in Table 4 remain near 1.0 for this period, but those in Table 5 rise well above unity. Since degrees of freedom decline going down the right side of Tables 4 and 5, the standard errors of estimate rise rather sharply, so that these coefficients merit less confidence than the earlier ones. If they are believed, however, they sug-

gest that for a time after 1965 interest rates actually overadjusted to changes in expectations. This result is not entirely implausible, for the strong upsurge in inflation beginning in 1965 may have awakened people to the importance of inflation for interest rates and caused them to overadjust to attempt to make up for past incomplete adjustments to expectations changes.

Since the estimates in Tables 4 and 5 do not uniformly rise as less pre-1965 experience is included in the estimation period, it appears that the increased responsiveness of interest rates to expectations in the 1960's is not due entirely to this later experience. Table 4 reveals a peak in estimates of  $a_1$  for the 1960:6 to 1970:12 period at 0.933, from which they decline substantially for periods beginning a few years later. Table 5 exhibits a similar pat-

TABLE 5—INTEREST RATES AND 12-MONTH EXPECTED RATES OF INFLATION FOR VARIOUS SUBPERIODS, 1952–1970  
 $i = a_0 + a_1 \dot{p}_{t+12}^*$   
 $i$  = 3-month Treasury Bill Rate

t	Subperiod 1952:6 through t			Subperiod t+6 through 1970:12		
	$a_1$	Standard Error of $a_1$	$R^2$	$a_1$	Standard Error of $a_1$	$R^2$
1957:6	.399	.2265	.173	1.082	.1149	.771
1957:12	.429	.2264	.190	1.122	.1174	.783
1958:6	.382	.2474	.104	1.063	.1174	.772
1958:12	.411	.2283	.147	1.059	.1221	.763
1959:6	.454	.2249	.180	1.087	.1269	.767
1959:12	.611	.2256	.297	1.108	.1246	.788
1960:6	.590	.2137	.292	1.092	.1297	.778
1960:12	.557	.2045	.274	1.061	.1320	.770
1961:6	.529	.1951	.261	1.025	.1320	.765
1961:12	.508	.1832	.261	.992	.1307	.769
1962:6	.507	.1750	.270	.960	.1351	.756
1962:12	.503	.1650	.284	.923	.1328	.759
1963:6	.512	.1612	.292	.900	.1459	.726
1963:12	.533	.1611	.302	.911	.1637	.697
1964:6	.555	.1568	.324	.889	.1816	.657
1964:12	.585	.1567	.341	.911	.2103	.618
1965:6	.600	.1593	.337	1.063	.2719	.588
1965:12	.645	.1617	.356	1.497	.3227	.695
1966:6	.680	.1466	.423	1.519	.3712	.663

Source: Expected prices from Livingston; observed consumer price index and interest rates from *Federal Reserve Bulletin*.

tern except that there is also a slightly higher peak for the 1958:6–1970:12 period. It thus appears that an important shift took place at the end of the 1950's, probably at the very end.

While we did not necessarily expect this shift in the relationship over time, it is consistent with the findings of another study of expectations. Yohe and Karnosky, when relating interest rates to weighted averages of past prices, found that the coefficients of past prices were higher for the period 1961–69 than for 1952–60. They concluded that price expectations had stronger effects on interest rates after 1960 than before. This conclusion is only suggested by their results, however, because their low coefficients for past prices before 1961 could hold because expectations were little influenced by actual price movements in that period. Yohe and

Karnosky neglect this very real possibility. Although we cannot reject completely this latter explanation, it does appear that expectations have in fact had a stronger effect on interest rates since 1959. To the extent that the measure of market expectations employed here is accurate—and there is no particular reason for assuming it less accurate before 1959 than after—it appears that expectations have had a stronger influence after 1959. That is, it is not that expectations were slower to adjust to actual price experience before 1960 than after. Rather, once formed, expectations had substantially less than their theoretically expected effects on interest rates in the earlier period.

This is a curious result, made more so by the arithmetic of the effects of interest rate changes on bond prices. The lower the level of market interest rates, the larger

the effect on the price of bonds of a given change in the expected rate of inflation. Since interest rates were lower in the 1950's than in the 1960's, this effect should have produced greater responsiveness in the earlier period. One explanation is that after 1952 people's consciousness was still dominated by the experience before the Treasury-Federal Reserve Accord of 1951, a period when interest rates were pegged and did not adjust for anything. Several years might have been required to change this neglect. It also remains conceptually possible that price expectations and the real rate were negatively related before 1959 but not after. It is difficult to imagine why this might have happened, but it cannot be ruled out completely.

Alternatively, it may be that transactions and information costs are such that at low rates of inflation it simply does not pay investors to bother adjusting to inflationary expectations. Some amount of effort is required to predict price behavior, including not only observing current prices but also obtaining information on current and projected values of other variables which are believed to influence future prices. If one determines that whatever the rate of inflation is, it is unlikely to be large, he might not bother looking into actual and projected values of these variables. Since the actual rate of inflation was substantially higher after 1959 than before, it may in fact have been not worth the bother for market participants to predict the future rate in the earlier period while it was in the later period. In such a case, the expectations generated before 1959 would be less accurate than those after, so that rational market participants would have given them less weight before 1959.

The above argument says that the cost of information on inflation is independent of the amount of inflation to be predicted, so that there are scale economies of information gathering as the expected rate in-

creases. On the other hand, the entire information-gathering cost function may have shifted downward around 1959, making it cheaper than before to obtain any amount of information for predicting inflation. If for either reason information costs were higher before 1959, market participants would have assembled less information in forming price expectations and thus would have put less faith in these estimates. Thus they probably would have discounted these expectations in interest rate calculations. Changes in information costs might have taken a number of different forms, including such items as better press coverage of economic news, greater theoretical economic expertise of market participants, better efforts of the government to assemble and disseminate economic measures, and the like.

If it were true that either the scale of inflation before 1959 made predictions greatly less profitable or that information costs were higher before 1959 than after, the effect should appear in the accuracy of price expectations. Information which is cheaper for either reason should be accompanied by forecasts which are closer to the mark, for the market has more incentive to assemble more information in forming expectations. Table 6 shows that survey respondents did indeed predict prices better after 1959 than before. Actual prices are regressed on their earlier predictions for the entire period and for the pre- and post-1959 subperiods. For both 6- and 12-month predictions the coefficients are closer to 1.0 and show statistical significance at a level closer to 0 in the later subperiod. When standard errors of the regressions are compared between subperiods, they are 25 and 37 percent lower for the 6- and 12-month predictions, respectively, in the post-1959 period. The coefficients were confirmed to be significantly different by a Chow test. For both the 6- and 12-month predicted price level the difference in the coefficients

TABLE 6—EXPECTED AND ACTUAL PRICE LEVELS, 1952-1970  
 $P_t = b_0 + b_1 P_{t-n}^*$ 

Period	$b_0$	$b_1$	$R^2$	<i>S.E.</i>
$n=6$				
1952:6-1970:12	0.929 (0.975)	0.99653 <sup>a</sup> (0.00912)	0.997	0.68964
1952:6-1959:12	3.017 (5.184)	0.97676 <sup>a</sup> (0.05429)	0.956	0.75384
1960:6-1970:12	-2.377 (1.325)	1.02423 <sup>a</sup> (0.01156)	0.997	0.55627
$n=12$				
1952:6-1970:12	6.047 <sup>a</sup> (6.974)	0.94139 <sup>a</sup> (0.00904)	0.997	0.72366
1952:6-1959:12	8.664 (5.715)	0.91536 <sup>a</sup> (0.05970)	0.940	0.87774
1960:6-1970:12	3.960 <sup>a</sup> (1.302)	0.95870 <sup>a</sup> (0.01125)	0.997	0.57790

<sup>a</sup> Denotes significant at the 0.05 level $P$  = Actual Consumer Price Index, from *Federal Reserve Bulletin* $P^*$  = expected price level, from Livingston

was significant well beyond the 0.01 level. Thus, relevant information costs seem to have decreased around 1959. Expectations were less accurate before 1959 than after, and it appears that the market accordingly put less faith in them in establishing market interest rates.

These findings also bear on another possible explanation of low coefficients of past price changes following from ideas put forth by Maurice Allais<sup>6</sup> and Milton Friedman and Anna Schwartz. Namely, the adjustment rate of expectations to changes in economic behavior may be influenced by the rate at which the latter changes are taking place. It could be that for some reason people adjust price expectations at a faster rate the higher the rate at which prices are changing. Thus, when the rate of inflation increases to a "high" rate the expected rate of inflation increases for two

reasons: 1) at a constant rate of adjustment of expected to actual inflation rates, the expected rate rises because the actual rate rises, and 2) the coefficient of adjustment of expected to actual rises, raising the portion of the actual rate incorporated with the expected rate. According to Allais' approach such coefficients are constant measured in terms of psychological time; i.e. a period in which some set of events takes place, such as an accumulated increase in prices of 10 percent. But actual behavior is measured in terms of calendar time, so that the observed speed of adjustment varies. Whereas Yohe and Karnosky advanced this as an explanation for their conclusion that expectations are slow to affect interest rates, it is in fact an explanation of why past price changes might be slow to affect expectations. The results here, however, suggest that this explanation is not relevant for explaining the low coefficients of past prices for 1952-59. For once formed, expectations had a small effect on interest

<sup>6</sup> Allais' best descriptions of the psychological time approach seem still to be in French and unavailable. His 1966 article does, however, refer to and attempt to employ this concept.

rates, a result which is separate from the speed at which expectations adjust to actual price experience.

### V. Summary and Conclusions

Although from 1897 to 1930 Fisher presented substantial evidence on the effect, the influence of price expectations on market interest rates in the United States<sup>6</sup> remained the subject of some doubts until recently when several studies documented its influence for U.S. data. These studies relied, however, on the hypothesis that the expected rate of inflation is dependent upon past rates of price change, so that this hypothesis was being tested along with Fisher's. The present study alleviates the need to test two relationships at once by using directly observed data on price expectations of an important segment of the market determining interest rates. When used for this purpose, these data reveal that market rates have been very strongly affected by expectations, particularly since 1959. A one percentage point 6- or 12-month expected rate of inflation is associated with interest rates which are higher by about a full percentage point for maturities of a year or less and by somewhat less for longer maturities. The results lend support to the hypothesis that the real rate of interest is not affected by price expectations over a six-month period and that interest rates fully adjust to expectations within six months. The estimates tend to support the hypothesis that expectations of a given term have less influence on yields as the term to maturity increases beyond the term of the expectation.

It also appears that expectations have had a much stronger effect on interest rates after 1959 than before. Since predictions of prices have been much more accurate

since 1959, it appears that information costs made predicting inflation less rewarding for the market before 1959. This result is consistent with one or both of the following hypotheses. First, actual inflation before 1959 may have been so small as to suggest that the benefits to be gained from predicting it would be less worth the trouble, whereas after 1959 the inflation rate increased so as to make predictions more worthwhile. Second, the entire function relating costs to the amount of information gathered may have shifted downward after 1959. For either reason, the accuracy of expectations seems to have increased since 1959, causing increased weight to be put on them by the market.

Although somewhat less conclusive, the results here also suggest that there may have been a particularly sharp change in the response of interest rates to price expectations in the post-1965 period of high continuing actual inflation. There is some evidence that interest rates actually over-adjusted to expectations for a time beginning in 1966, possibly in an effort to compensate for earlier incomplete adjustment.

### APPENDIX

#### *Listings of Livingston survey respondents, 1947 and 1970*

June 1947:

Charles O. Hardy, adviser to the Joint Committee on the Economic Report; C. A. Sienkiewicz, president, Central-Penn National Bank, Philadelphia; Rufus Tucker, General Motors; Frank L. Valenta, vice president, Distributors Group, Inc.; Robert R. Nathan of Robert R. Nathan Associates; J. Frederic Dewhurst, Twentieth Century Fund and author of the recent monumental work, "America's Needs and Resources"; Boris Shiskin, American Federation of Labor.

Also, E. G. Bennion, Standard Oil Co. (N.J.); Alan H. Temple, vice president, National City Bank of New York; Woodlief Thomas, director of research and statistics,

<sup>6</sup> Philip Cagan in 1956 provided very strong evidence of price-expectations effects in postwar European inflations.

Federal Reserve Board; James F. Hughes, Auchincloss, Parker & Redpath, members of the New York Stock Exchange; Herbert Stein, economist for the Committee for Economic Development, and winner of the \$25,000 Pabst award; Donald R. Smith, partner, Scudder, Stevens & Clark, investment counsel, Joseph K. Heyman, Joseph K. Heyman Co., Atlanta; Morris A. Copeland, Federal Reserve Board; Dexter M. Keezer, McGraw-Hill Publishing Co.; Charles D. Stewart, U.S. Bureau of Labor Statistics; Robert C. Shook, vice president, International Statistical Bureau; Donald F. Bishop, investment counsel, Philadelphia; Solomon Barkin, director of research, Textile Workers Union of America; O. C. Stine, U.S. Department of Agriculture; Arthur R. Upgren, University of Minnesota; Robert Coltman, vice president, Provident Trust Co., Philadelphia; Nathaniel R. Whitney, Procter & Gamble Co.; John Patterson, National Shoe Manufacturers Ass.; Lazare Teper, International Ladies' Garment Workers Union; F. E. Richier, General Foods Corp.; Wilson Wright, Armstrong Cork Co.; Glen G. Munn, Paine, Webber, Jackson & Curtis, members of the New York Stock Exchange; Adolf G. Abramson, SKF Industries.

December 1970:

Gardner Ackley, Henry Carter Adams Professor, University of Michigan; Daniel S. Ahearn, vice president, Wellington Management Co.; Louis H. Bean, Arlington, Va.; Edward G. Boehne, Federal Reserve Bank of Philadelphia; Karl Brandt, emeritus professor, Stanford University.

Ewan Clague, Washington, D.C.; Carrol L. Coburn, director of research, United Auto Workers; Philip E. Coldwell, president, Federal Reserve Bank of Dallas; George W. Coleman, American Bankers Association; Andrew T. Court, Detroit, Mich.; James M. Dawson, vice president, National City Bank of Cleveland; John V. Deaver, manager, economics department, Ford Motor Co.; Robert J. Eggert, staff vice president, RCA Corp.; Richard W. Everett, vice president, Chase Manhattan Bank.

Phillip John Fitzgerald, Dean Witter & Co.; William C. Freund, vice president, New York Stock Exchange; Douglas Greenwald, McGraw-Hill, Inc.; David L. Grave, International Business Machines Corp.; A. Gilbert Heebner, senior vice president, Philadelphia National Bank; Peter Henle, Bureau of Labor Statistics; Homer Jones, vice president, Federal Reserve Bank of St. Louis; Vernon E. Jirikowic, research director, International Association of Machinists and Aerospace Workers.

J. W. Kendrick, American Airlines, Inc.; L. R. Klein, Wharton School, University of Pennsylvania; Richard W. Lambourne, vice president, Crocker-Citizens National Bank; Robert J. Landry, Dun & Bradstreet, Inc.; Werner Lehnberg, Goodbody & Co.; Wesley Lindow, senior vice president and secretary, Irving Trust Co.; Carl H. Madden, U.S. Chamber of Commerce; Edmund A. Mennis, senior vice president, Republic National Bank of Dallas; H. LeBrec Micoleau, General Motors Corp.

Robert R. Nathan, Robert R. Nathan Associates, Washington, D.C.; Robert J. Oster, Bank of America; Louis J. Paradiso, senior associate, Tyson, Belzer, and Associates, Inc.; Sanford F. Parker, Fortune Magazine; Robert W. Paterson, School of Business and Public Administration, University of Missouri; W. H. Peterson, United States Steel Corp.; Francis H. Schott, vice president, Equitable Life Assurance Society; Irving Schweiger, Graduate School of Business, University of Chicago; Beryl W. Sprinkel, vice president, Harris Trust & Savings Bank, Chicago.

Lazare Teper, director of research, International Ladies Garment Workers Union; W. W. Tongue, University of Illinois; M. L. Upchurch, U.S. Department of Agriculture; Henry C. Wallich, Yale University; Robert L. Weidenhammer, Rockville, Md.; Rudolph L. Weissman, W. E. Hutton & Co.; Gary M. Wenglowksi, Goldman, Sachs & Co.; J. P. Wernette, University of Michigan; Simor N. Whitney, New York University; Hans A. Widenmann, Loeb, Rhoades & Co.; Seymour

Wolbein, Dean, School of Business Administration, Temple University.

Sources: 1947: *Washington Post*; 1970: *The Sunday Bulletin*.

#### REFERENCES

- M. Allais, "A Restatement of the Quantity Theory of Money," *Amer. Econ. Rev.*, Dec. 1966, 56, 1123-57.
- P. Cagan, "The Monetary Dynamics of Hyperinflation," in M. Friedman, ed., *Studies in the Quantity Theory of Money*, Chicago 1956, 23-117.
- I. Fisher, *The Theory of Interest*, New York 1930.
- M. Friedman and A. J. Schwartz, "Trends in Money, Income, and Prices 1867-1966," unpublished study, Nat. Bur. Econ. Res. 1967, ch. 4.
- W. E. Gibson, "Price-expectations Effects on Interest Rates," *J. Finance*, Mar. 1970, 25, 19-34.
- R. A. Kessel and A. A. Alchian, "Effects of Inflation," *J. Polit. Econ.*, Dec. 1962, 70, 521-37.
- J. A. Livingston, "Business Outlook," *The Washington Post*, July 6, 1947.
- R. A. Mundell, "Inflation and Real Interest," *J. Polit. Econ.*, June 1963, 71, 280-83.
- M. Porter, "Anticipated Inflation and the Real Rate of Interest," unpublished manuscript.
- T. J. Sargent, "Commodity Price Expectations and the Interest Rate," *Quart. J. Econ.*, Feb. 1969, 83, 127-40.
- S. Turnovsky, "Empirical Evidence on the Formation of Price Expectations," *J. Amer. Statist. Ass.*, Dec. 1970, 65, 1441-54.
- W. P. Yohe and D. S. Karnosky, "Interest Rates and Price Level Changes," *Fed. Reserve Bank St. Louis Rev.*, Dec. 1969, 51, 19-36.
- Federal Reserve Bulletin*, various issues, 1952-71.
- Washington Post*, July 6, 1947.
- The Sunday Bulletin* (Philadelphia, Pa.), Dec. 27, 1970.

# Capital Gains and the Aggregate Consumption Function

By KUL B. BHATIA\*

The purpose of this paper is to explore the effect of capital gains on aggregate consumption. Capital gains have been an important source of income in the United States in recent years. During 1947-64, about \$150 billion of realized gains were reported for tax purposes, but accrued gains (whether realized or not) have been much larger and amounted to about 20 percent of personal income in some years.<sup>1</sup> Most time-series studies of the aggregate consumption function in the United States have relied on the official estimates of income and saving which do not include capital gains; consequently, capital gains have been generally ignored.<sup>2</sup> What is puzzling, however, is that in one or two instances where such gains were included (for example, see John Arena (1963, 1965)), it was found that they had no significant influence on aggregate consumption or consumer spending. This result is explained by pointing out that capital gains accrue mainly to upper income groups, are mostly "transitory" in character, and are not treated as income be-

cause accrued gains may never be realized and spent if asset prices decline.

These arguments, by and large, are inconclusive. Gains do not have to be realized to be spent; savings in other forms can be reduced, or one can borrow on the security of the appreciated value of one's asset holdings. Not all gains are "windfalls," but even if they were, there is some empirical evidence to show that marginal propensity to consume out of windfalls need not always be zero,<sup>3</sup> and the effect of transitory income on consumption depends on the source of income variation and may not be negligible in all cases.<sup>4</sup>

The burden of the present paper is that capital gains have not been studied in an appropriate theoretical model in the existing literature, nor have any direct estimates of gains been used in empirical estimation.<sup>5</sup> We incorporate capital gains into the permanent income framework and test several hypotheses about the effect of capital gains on aggregate consumption in the light of new and, in many ways, more consistent estimates of accrued capital gains derived by Bhatia (1969). To anticipate the conclusions of the paper, when an appropriate distributed lag function is specified, both accrued and realized capital gains turn out to be significant variables

\* University of Western Ontario. An earlier version of the paper was presented to a meeting of the Econometric Society. I would like to thank Zvi Griliches, Scott Clark, and Tom Valentine for many stimulating discussions. Russell Boyer and Joel Fried commented on an earlier version of the paper, and Chai-Yan Kuo helped with computations. The research on the paper was supported by a grant from the Canada Council.

<sup>1</sup> Accrued gains represent the true change in the economic position of individuals and would be included in comprehensive income concepts like those proposed by Robert Haig, J. R. Hicks, and Henry C. Smons.

<sup>2</sup> The inclusion of capital gains in national income has been a debatable topic for a long time. For a discussion of the controversy, see Bhatia (1970) and Michael McElroy.

<sup>3</sup> Ronald Bodkin concludes that "the true (population) value of the marginal propensity to consume out of windfall income is, in all likelihood, quite high" (p. 613). Also see the references cited there, especially Lawrence Klein and Nissan Liviatan.

<sup>4</sup> Jacob Viner shows that transitory income caused by short-run changes in employment significantly affects consumption.

<sup>5</sup> Arena (1965) is an exception, but only stock market gains are used in that study.

in the aggregate consumption function.

The theoretical structure for the paper is developed in Section I, Section II contains a description of capital gains during 1948-64, the time period covered by this study, the empirical estimates are presented in Section III, and Section IV deals with the policy implications of our results. For brevity, the term "gains" is used to denote both capital gains and losses.

### I. Theoretical Considerations

Whenever wealth or consumer net worth is used to explain consumer behavior, capital gains, which reflect changes in the value of assets held by consumers, are implicitly included in the consumption function. Thus, in the "life-cycle" hypothesis of saving suggested by Franco Modigliani and Richard Brumberg where consumption is proportional to the present value of total resources that would accrue to an individual during the rest of his life, and in studies by Alan Spiro, and Robert Ball and Pamela Drake in which households' net worth is used as an explanatory variable, capital gains are incorporated.<sup>6</sup> Because no suitable estimates of capital gains were available, Ball and Drake excluded gains from their formal analysis, and Spiro included corporate retained earnings—obviously a proxy for capital gains—in income.

Arena (1963) postulates consumption as a function of the gap between desired target wealth ( $W_t^*$ ) and potential target wealth, the latter consisting of initial wealth ( $A_{t-1}$ ) and expected capital gains

( $G_t^e$ ). Arena assumes that expected gains are a linear function of current period gains ( $G_t^e = a_1 + a_2 G_t$ ), and derives a consumption function of the following form:

$$(1) \quad C_t = \beta_1 + \beta_2 Y_t + \beta_3 A_{t-1} + \beta_4 G_t$$

where  $Y_t$  is current disposable income, and capital gains now appear as an explicit variable in the function. Arena found that  $\beta_4$  was insignificant, and  $\beta_3$  and  $\beta_4$  did not differ significantly.

It is not correct to conclude from these results that capital gains have no significant affect on consumption because only current period gains appear in equation (1). Given the nature of capital gains, it is likely that people would revise their estimates of wealth after gains have accrued for some time, and consumption would respond with a lag longer than one period. These considerations can be easily incorporated into the analysis in terms of the permanent income framework which we shall adopt to formulate some hypotheses about the influence of capital gains on aggregate consumption.

### The Model

Following the simplest version of the permanent income hypothesis, we postulate that consumption is a constant function of permanent income:

$$(2) \quad C_t = k Y_t^p + u_t$$

$C_t - u_t$  can be interpreted as planned consumption. Although capital gains have been treated as a part of wealth in almost all the earlier studies in this area, many income theorists would treat capital gains like any other income and include them directly in a comprehensive income concept. Therefore, it is useful to distinguish between the "income" and the "wealth" approach to study the effect of capital gains on aggregate consumption. Within the framework of our model, the main difference between the two approaches will

<sup>6</sup> Albert Ando and Modigliani fit the following equation:

$$C_t = \alpha_1 Y_t^e + \alpha_2 Y_t + \alpha_3 A_{t-1}$$

where  $C_t$ ,  $Y_t^e$ ,  $Y_t$  and  $A_{t-1}$  represent aggregate consumption, current nonproperty income, expected annual nonproperty income, and net worth, respectively. However, as Arena (1964) points out, in their computations, net worth for the period is based on the *average price* during the year; therefore, it includes part of the gains accruing in that year.

be in the specification of permanent income.

### *The Income Approach*

If gains are included in income directly, permanent income will be a linear function of expected disposable income ( $Y_t^e$ ) and expected capital gains ( $G_t^e$ ). Equation (2) can then be rewritten as

$$(3) \quad C_t = k_1 Y_t^e + k_2 G_t^e + u_t$$

$Y_t^e$  and  $G_t^e$  can be related to observed variables by assuming that<sup>7</sup>

$$Y_t^e - Y_{t-1}^e = (1 - \alpha)(Y_t^d - Y_{t-1}^e), \\ 0 < \alpha < 1$$

where  $Y_t^d$  is disposable income. This hypothesis about the formation of expectations implies that expected income is a geometrically declining distributed lag function of past incomes, i.e.,

$$Y_t^e = (1 - \alpha) \sum \alpha^i Y_{t-i}^d$$

If we assume further that  $G_t^e = (1 - \alpha) \sum \alpha^i G_{t-i}$ , we can use a transformation described by Leen Koyck on equation (3) to derive

$$(4) \quad C_t = k_1(1 - \alpha) Y_t^d + k_2(1 - \alpha) G_t \\ + \alpha C_{t-1} + u_t - \alpha u_{t-1}$$

If capital gains affect consumption, their coefficient should be positive and significant. Furthermore, if capital gains are like other types of income,  $k_1$  should equal  $k_2$  in equation (4). This suggests our first two hypotheses:

**H1**  $k_2$  is positive and significant, and

**H2**  $k_2 = k_1$

### *The Wealth Approach*

If capital gains affect consumption via wealth, we can assume that permanent

income is the sum of expected labor income ( $Y_t^w$ ) and expected income from wealth,

$$(5) \quad Y_t^p = Y_t^w + \beta W_t^*$$

where  $\beta$  is the "normal return" on wealth, and  $W_t^*$  is the expected wealth relevant for making consumption decisions in period  $t$ . By substituting (5) into (2) we get:

$$(6) \quad C_t = k Y_t^w + k\beta W_t^*$$

We further assume that the current level of expected wealth is the sum of actual wealth at the end of the previous period ( $W_{t-1}$ ), and a lagged function of accrued capital gains,

$$(7) \quad W_t^* = W_{t-1} + W(L)G_t$$

where  $W(L)$  is a lag-generating function.

Equation (7) implies that people do not feel richer or poorer with every movement in asset prices. However, some sort of an average of past accrued gains gets incorporated into their estimation of wealth and affects consumption. The term  $W_t^*$ , thus, is a subjective value concept which differs from the actual market value of wealth inasmuch as it includes *expected* gains whereas only *actual* gains would be included in  $W_t$ . Substituting equation (7) into (6) we get:

$$(8) \quad C_t = k Y_t^w + k\beta W_{t-1} + k\beta W(L)G_t$$

If our specification is valid, the gains' coefficient in (8) should be positive and significant. The third hypothesis, therefore, is

**H3**  $k\beta > 0$  and significant.

However, if  $k\beta$  is found to be not significantly different from zero, we can conclude that capital gains have no significant influence on aggregate consumption.

### *The Role of Realized Gains*

In a world of perfect capital markets and equal taxes on all types of income, a dis-

<sup>7</sup> This is the familiar adaptive expectations model in which expectations are revised in proportion to the error associated with the previous level of expectations.

inction between realized and accrued gains would be unnecessary. However, at present, borrowing on the security of some types of assets (for example, corporate stock) is restricted; so some people might believe that a bird in hand is better than two in the bush and treat realized gains as more "real" income than accrued gains. Besides, accrued gains are taxed more lightly, if at all, than realized gains; therefore, it is likely that the two types of gains might affect consumption differently. Tax considerations would lead to a smaller coefficient for realized gains, but other considerations might offset, and even reverse this tendency. Accordingly, we shall estimate equations (3) and (8) for both realized and accrued gains in Section III.<sup>8</sup>

<sup>8</sup> The possibility of an identification problem should be noted in this context because often a consumption

In the next section we discuss several aspects of capital gains accruing during 1948-64, the period covered by this study. Empirical estimates of the various equations are presented in Section III.

## II. Postwar Capital Gains

It is widely believed that capital gains are windfall income and accrue mostly to upper income groups, mainly on their holdings of corporate stock. Table 1 presents annual data on realized gains reported on individual income tax returns and gains accruing on four asset-types which are recognized as "capital assets"

plan might precede realization of accrued gain. For example, to buy a car, a person may sell some of the corporate stock he owns. In this case, realized gains would appear to affect consumption significantly, although, in fact, it was consumption which prompted the realization of a gain.

TABLE 1—ACCRUED AND REALIZED CAPITAL GAINS, 1948-64  
(billion dollars)

	Corporate Stock (1)	Nonfarm Residential Real Estate (2)	Farm Real Estate (3)	Livestock (4)	Total Accrued Gains (5)	Total Realized Gains (6)	Reported Personal Income (7)	Reported Personal Saving (8)
1948	- 0.65	11.89	3.11	1.2	15.55	4.20	210.2	13.4
1949	10.19	-12.41	-0.81	-1.7	- 4.73	3.01	207.2	9.4
1950	23.44	15.95	6.59	3.6	49.58	5.81	227.6	13.1
1951	16.07	16.80	8.34	1.4	42.61	6.02	255.6	17.3
1952	10.47	16.88	2.43	-5.3	24.48	4.86	272.5	18.1
1953	- 1.62	- 4.02	-1.41	-3.0	-10.05	4.00	288.2	18.3
1954	57.18	5.35	2.15	-0.8	63.88	6.66	290.1	16.4
1955	52.37	15.98	4.55	-0.6	72.30	9.33	310.9	15.8
1956	6.34	29.36	7.29	0.7	43.69	8.97	333.0	20.6
1957	-34.65	4.90	7.33	3.1	-19.32	6.93	351.1	20.7
1958	91.89	2.43	7.85	3.2	105.37	8.58	361.2	22.3
1959	35.05	15.63	5.77	-2.9	53.55	12.33	383.5	19.1
1960	-14.40	17.51	3.61	0.3	7.02	10.38	401.0	17.0
1961	81.38	18.70	4.92	0.4	105.40	16.12	416.8	21.2
1962	-52.95	- 7.15	6.34	0.3	-53.46	11.01	442.6	21.6
1963	73.91	7.85	8.75	-1.9	88.61	12.85	465.5	19.9
1964	59.07	30.42	9.29	-1.2	97.58	15.71	497.5	26.2
Totals	413.09	186.07	86.10	-3.2	682.06	146.77		310.4

Source: Cols. 1-4: See Bhatia (1970).

Col. 5: Sum of cols. 1-4.

Col. 6: Derived from *The Statistics of Income, Individual Income Tax Returns*, various years.

Cols. 7-8: *Economic Report of the President*, Jan. 1969, Table B-15, p. 244.

for purposes of income tax.<sup>9</sup> The estimates show that during 1948-64, accrued net capital gains on corporate stock, real estate, and livestock owned by individuals amounted to \$682.46 billion which implies that gains accrued at an average rate of about \$40 billion a year. Corporate stock accounted for more than 60 percent of all accrued gains during this period but in many years, especially before 1955, gains and losses on real estate (columns 2 and 3 in Table 1) exceeded those on corporate stock. Realized gains, amounting to \$146.77 billion, are much smaller than accrued gains. It is difficult, however, to estimate precisely what proportion of gains accruing during these years was actually realized.<sup>10</sup>

To get an idea of relative magnitudes, personal income and personal saving estimated in the U.S. National Accounts is also reported in Table 1. It is clear that realized gains have rarely amounted to 3 or 4 percent of personal income but in many years, accrued gains have been as much as 25 percent of personal income. For the entire period, realized gains are about 47 percent of personal saving but accrued gains are much larger and amount to more than twice the official estimate of personal saving.

#### *Income Distribution of Capital Gains*

Tables 2 and 3 present some data on the income distribution of realized gains for a

TABLE 2—REALIZED CAPITAL GAINS AS PERCENT OF REALIZED INCOME<sup>a</sup>

Income Bracket (dollars in thousands)	1950	1955	1960	1964
Below 25	3.2	2.9	2.0	1.9
25-50	9.1	12.6	8.8	10.5
50-100	13.4	19.9	17.1	17.9
100 and above	37.9	49.1	54.2	49.1

<sup>a</sup> Based on data reported in *The Statistics of Income, Individual Income Tax Returns*, various years. Realized income is derived by including all realized gains in adjusted gross income.

few years. It is apparent from Table 2 that the upper income groups realize large portions of their income in capital gains. However, Table 3 shows that individuals with incomes below \$25,000 account for more than 40 percent of all gains realized in a given year. This is due to the relatively large number of individuals in this income category.

Similar statistics cannot be presented for accrued gains because not much is known about their income distribution. However, for the few years for which attempts have been made to estimate their income distribution, the results suggest that the top 5 percent of income recipients account for about 50 percent of all accrued gains, and less than 20 percent of all gains

TABLE 3—PERCENT OF ALL CAPITAL GAINS REALIZED BY VARIOUS INCOME BRACKETS<sup>a</sup>

Income Bracket (dollars in thousands)	1950	1955	1960	1964
Under 25	50.4	47.6	47.6	41.2
25-50	12.4	15.0	13.0	16.0
50-100	10.5	12.2	12.0	13.1
100 and above	26.7	24.9	27.4	29.7

<sup>a</sup> Based on data reported in *Statistics of Income, Individual Income Tax Returns*, various years.

<sup>9</sup> Throughout this paper we shall follow the definitions used in the Internal Revenue Code. Thus we define capital gains as nominal gains without any adjustment for changes in the general price level. Although many assets are treated as capital assets for tax purposes, the four asset categories on which data are presented in Table 1 account for most of the gains accruing to the household sector. It is difficult to make comparable estimates of accrued gains for other types of assets due to inadequate data.

<sup>10</sup> There are many reasons for this statement. Ignoring the problem of under-reporting of realized gains, the estimates of accrued gains most likely provide a lower bound because all asset categories could not be covered. Besides, some of the gains realized during 1948-64 could have accrued before 1948.

accrue to the lowest three income quintiles.<sup>11</sup>

### *Are Capital Gains "Transitory" Income?*

Transitory income is defined by Milton Friedman to include income due to all factors "... that are likely to be treated by the unit affected as 'accidental' or 'chance' occurrences, though they may, from another point of view, be the predictable effect of specifiable forces.... In statistical data, the transitory component includes also chance errors of measurement" (p. 22). In terms of this definition, it is very difficult to decide whether capital gains are transitory income because the data do not give us much information on how individuals treat such gains. Some light on the matter, however, could be shed by statistics on the duration of asset holding, i.e., the time period for which an asset is held before gains are realized.

In 1962—the only year for which such data are available—about 5 percent of gross gain on corporate stock was realized on stock held for less than six months, about 6 percent on that held between six months and a year, and about 20 percent on corporate stock held for five-to-ten years. The number of long-term transactions was more than twice that of short-term transactions.<sup>12</sup> Real estate assets owned for a year or more accounted for approximately 75 percent of gross gain realized on such assets in 1962, and about 50 percent of the gain was realized on real estate held for ten years or more. It is clear

that a sizable amount of the gains realized in 1962 had accrued over several past years. Since the observation relates to only one year which also happened to be an abnormal year, at least for corporate stock, no definite conclusions can be derived from it about the nature of capital gains. The evidence, however, does question the putative notion that all capital gains are transitory or windfall.

To summarize the discussion in this section, during 1948–64 accrued capital gains have been quite large relative to personal income and saving, realized gains have been much smaller than accruals, and might not always be transitory or windfall income. Although corporate stock accounts for the bulk of capital gains, accrued gains on other assets have not been negligible.

### III. Empirical Verification and Estimation

Empirical estimates of the model outlined in Section I are presented in this section. The period of observation is 1948 through 1964. Consumption,  $C_t$  (defined as personal expenditures on nondurable goods and services, plus depreciation on consumer durable goods); labor income net of taxes,  $Y_t^s$ ; personal disposable income,  $Y_t^d$ ; accrued gains,  $G_t$ ; realized gains,  $G_t^r$ ; and wealth,  $W_t$ , are all measured in billions of current dollars.<sup>13</sup>

#### *The Measurement of Expected Variables*

Equations (3) and (8) derived in Section I have several variables which cannot be directly observed. The income approach uses expected disposable income,  $Y_t^e$ : the

<sup>11</sup> Compare McElroy, ch. 5. Unfortunately there are no basic data on the income distribution of accrued gains and losses. These results are derived by piecing together information from several indirect sources and may not be very accurate. Also, the concepts used by McElroy are somewhat different from those in this paper. The income distribution of capital gains will be examined more thoroughly in my 1972 study.

<sup>12</sup> In income tax parlance, short-term transactions denote a sale within six months of the purchase, i.e., a holding period of less than six months. All other transactions are long term.

<sup>13</sup> I would like to thank George Craig, Northern Illinois University, for providing me with data on  $C_t$  and  $Y_t^d$ . These data can be found in his Ph.D. thesis and have been obtained by procedures described in Ando and Brown, Data Appendix. Data on  $Y_t^s$ ,  $G_t$ , and  $G_t^r$  are from Bhatia (1969). Unpublished data on consumer net worth were kindly provided by the Board of Governors of the Federal Reserve System in a letter to the author.

wealth approach relies on  $Y_t^w$ , expected nonproperty income; and  $G_t^e$  appears in both equations (3) and (8). Several hypotheses have been suggested in the literature about the formation of income expectations. Following Friedman, most studies of the permanent income have relied on the distributed lag approach, which implies that expected income is an exponentially weighted average of past incomes. Mincer defined  $Y_t^w$  as full employment labor income, and Ando and Modigliani defined average expected income as current income adjusted for a possible scale factor—the scale factor being substantially smaller for those currently unemployed than that for the fully employed. The only precedent for expected capital gains is Arena (1963) who defined expected gains simply as a linear function of current period gains.

We shall adopt the distributed lag approach for expected gains, because their effect on consumption is likely to be spread over more than one year. The various hypotheses about income expectations are theoretically plausible, but they all require arbitrary decisions on many points: for example, the weights in the distributed lag approach, or the definition and measurement of full employment labor income, etc. However, if we adopt the distributed lag approach for income also, the estimation becomes somewhat less complicated as illustrated by equation (4), and for this reason we shall assume that expectations about income and gains are based on past experience.

*Regression Results for the Income  
Approach: Equations  
(3) and (4)*

Estimates for equation (4), for both accrued and realized gains ( $G^r$ ), using the instrumental variables method suggested by Liviatan are as follows:

$$C_t = 0.434C_{t-1} + 0.525Y_t^d + 0.002G_t^e \quad (5.11) \quad (7.27) \quad (0.16)$$

$$R^2 = 0.999, SEE = 1.75$$

$$C_t = 0.458C_{t-1} + 0.497Y_t^d + 0.280G_t^r \quad (6.22) \quad (8.06) \quad (1.45)$$

$$R^2 = 0.999, SEE = 1.61$$

The  $t$ -values are enclosed in parentheses. In both cases  $Y_t^d$ ,  $Y_{t-1}^d$ , and  $G_t$  were used as the instrumental variables.<sup>14</sup>

All the coefficients have the expected signs (positive). The coefficients of disposable income and lagged consumption are highly significant in both cases, but the coefficient of accrued gains is very small and insignificant. We, therefore, reject **H1** and **H2**: given the validity of our model, accrued capital gains do not affect consumption via income.<sup>15</sup> The coefficient of realized gains, however, is significant at the 90 percent level and much larger than the coefficient of accrued gains.

It must be pointed out that the above conclusions hold only if expectations about income and gains are formed according to the same geometric lag function on which equation (4) is based. Thus, in rejecting **H1**, we could be rejecting the assumption about the equality of the two lag coefficients rather than the importance of accrued capital gains in the aggregate consumption function. In order to isolate these two tests, we let income and gains follow different distributed lag functions

<sup>14</sup> The estimation of (4) is complicated by the disturbance terms which are likely to be autocorrelated. Ordinary least squares, therefore, would yield biased, inconsistent, and inefficient estimates. Liviatan's procedure, although inefficient, is simple to apply and yields consistent estimates. The inefficiency, however, is not critical to the interpretation of the results because direct estimates of equation (4) allowing for different lag structures on income and gains and a well-behaved error term are presented later.

<sup>15</sup> This is the expected result if capital gains are treated as "transitory income" so that they are added to wealth in the first instance. See, for example, Michael Landsberger.

TABLE 4—ESTIMATES OF THE COEFFICIENTS OF THE CONSUMPTION FUNCTION, EQUATION (3)

	Constant	$Y_t^e$ ( $k_1$ )	$G_t^e$ ( $k_2$ )	SEE	$R^2$	D.W.
Accrued Gains	-7.62 (-3.20)	0.962 (123.0)	-0.007 (-0.72)	1.44	0.999	0.95
	—	0.953 (356.9)	-0.007 (-0.48)	2.03	0.999	0.97 <sup>a</sup>
Realized Gains	0.211 ( 0.02)	0.897 (17.31)	1.171 ( 1.15)	1.49	0.999	0.86
	—	0.899 (102.0)	1.159 ( 3.78)	1.43	0.999	0.85 <sup>a</sup>

Note: *t*-values are enclosed in parentheses. Expected income and gains follow different geometric lag functions. Effective number of observations in each case is 13.

<sup>a</sup> The D.W. statistic must not be calculated on the residuals of a regression equation without a constant term, but on those of an auxiliary regression on the same exogenous variables including a constant. Compare Edmond Malinvaud, p. 508. The D.W. statistics reported here and in other tables have been calculated in this manner.

for the results reported in Table 4.<sup>16</sup> Specifically, let

$$Y_t^e = (1 - \alpha) \sum_{i=0}^{\infty} \alpha^i Y_{t-i}^d, \quad \text{and}$$

$$G_t^e = (1 - \lambda) \sum_{i=0}^{\infty} \lambda^i G_{t-i}, \quad 0 < \alpha, \lambda < 1$$

This specification has the further advantage that we get away from the complicated error terms of equation (4). The estimation procedure, however, is now more intricate: it involves varying  $\alpha$  and  $\lambda$  over a predetermined range, calculating  $Y_t^e$  and  $G_t^e$ , estimating equation (3), and selecting those values of the regression coefficients for which the residual sum of squares is minimized.<sup>17</sup>

<sup>16</sup> We have retained the geometric lag form for ease of computations. Gains and income now have different variables of "adjustment"— $\lambda$  and  $\alpha$ , respectively.  $Y_t^e$  and  $G_t^e$  are weighted moving averages of past incomes and gains, respectively—weights adding up to unity in each case. This is similar to the formulation used by Friedman.

<sup>17</sup> If the model is correctly specified and the search procedure adequate, these estimates would be maximum likelihood estimates. See Zvi Griliches. The computations were done by an iterative program, SCAN, written

Once again, accrued gains are not significant in the income approach, but the coefficient of realized gains is greater than zero at the 95 percent level of significance. The most important figure reported in Table 4, however, is the Durbin-Watson statistic which falls considerably short of 2 in all cases, indicating a marked positive serial correlation in the residuals of equation (3). Consequently, although the coefficients would be unbiased, their reliability is in serious doubt. The serial correlation, therefore, should be corrected before interpreting these results.

In Table 5, we report estimates of equations (3) derived by using the Cochrane-Orcutt procedure.<sup>18</sup> The results change but

by the author. For the results reported in Table 4, we used  $\alpha = (1 - B + A)$ , where  $B$  is the initial period weight, and  $A$  a trend variable (equal to 0.03). For details of this procedure, see Colin Wright.  $\alpha$  and  $\lambda$  were allowed to vary from 0.1 to 0.99. Because of the small number of observations, we computed only 5 yearly moving averages. The weights added to unity in all cases.

<sup>18</sup> The procedure uses an ordinary least square regression to form an initial guess of  $\rho$ , the first-order serial coefficient. All data are transformed by  $\rho$  (e.g.:  $X_t - \rho X_{t-1}$ ), regression is run on transformed data, a new  $\rho$  is estimated and the process continues. Compare Donald Cochrane and Orcutt (1949).

TABLE 5—ESTIMATES OF THE COEFFICIENTS OF THE  
CONSUMPTION FUNCTION, EQUATION (3)  
Cochrane-Orcutt Estimates

	Constant	$Y_t^e$	$G_t^e$	$\rho$	SEE	$R^2$	D.W.
Accrued Gains	-10.96 (-3.52)	0.971 (102.07)	-0.004 (-0.72)	0.38 (1.41)	1.12	0.999	0.74
	—	0.953 (228.53)	-0.005 (-0.67)	0.69 (3.29)	1.53	0.999	1.11
Realized Gains	-6.038 (-0.855)	0.931 (22.25)	0.599 (.076)	0.50 (2.01)	1.18	0.999	0.88
	—	0.896 (62.02)	1.21 (2.49)	0.52 (2.12)	1.16	0.999	0.87

Note:  $t$ -values are in parentheses. Effective number of observations in each case is 12.

little: the  $R^2$  remains extremely high, the coefficients have roughly the same magnitudes and  $t$ -values as before, and the Durbin-Watson statistic is still very low. Consequently, our interpretation of the regression coefficients does not change. The inability of the Cochrane-Orcutt technique to correct the serial correlation in equation (3) suggests that either (3) is misspecified, or the error terms do not follow a first-order auto-regressive scheme as assumed by the adjustment procedure. Both  $G_t^e$  and  $Y_t^e$  in (3) have been computed as five yearly moving averages, so it is possible that  $u_t$  might be significantly correlated with  $u_{t-2}$  and still earlier disturbance terms. To test this possibility, we report below the estimates of equation (3), after adjusting for second-order serial correlation.

The coefficients of first- and second-order serial correlation— $\rho_1$ , and  $\rho_2$ , respectively—are significant in all cases. The Durbin-Watson statistic suggests that errors are now serially independent. Also, the standard errors of estimate are lower than those reported in Table 5. All except the coefficient of realized gains remain virtually unchanged. Expected income is the most significant variable, with a marginal propensity to consume about 0.9 or more. The coefficient of accrued gains is still negative and insignificant, but that of

realized gains is significantly greater than zero at the 90 percent level when the constant term, which is insignificant in case of realized gains, is dropped from the equation. Further, we cannot reject the hypothesis that the coefficient of expected realized gains is equal to that of expected income.<sup>19</sup> However, not much should be read into this result because the coefficient of realized gains is quite unstable: it becomes negative and insignificant when a constant term is included in the equation.

#### *Regression Results for the Wealth Approach*

The exogenous variables in equation (8) are net worth ( $W_{t-1}$ ), expected labor income ( $Y_t^w$ ), and expected capital gains [ $W(L)G_t$ ]. As in the income approach, we assume that  $Y_t^w$  and  $W(L)G_t$  are approximated by geometric lag functions of past labor incomes and gains, respectively. Let

$$Y_t^w = (1 - \alpha) \sum_{i=0}^{\infty} \alpha^i Y_{t-i}^s$$

and

$$W(L)G_t = (1 - \lambda) \sum_{i=0}^{\infty} \lambda^i G_{t-i}$$

where  $Y_t^s$  represents labor income and  $G_t$  is

<sup>19</sup> The  $t$ -value for the null-hypothesis that the two coefficients are equal is very small.

TABLE 6—ESTIMATES OF THE CONSUMPTION FUNCTION, EQUATION (3)  
Correction For Second-Order Serial Correlation

	Constant	$Y_t^e$	$G_t^e$	$D.W.$	$SEE$	$\rho_1$	$\rho_2$
Accrued Gains	-3.718 (-2.56)	0.959 (117.8)	-0.003 (-1.19)	1.86	0.74	0.90 (4.1)	-0.44 (-2.4)
Realized Gains	-4.16 (-1.59)	0.955 ( 33.83)	-0.046 (-0.08)	1.48	0.86	0.87 (3.1)	-0.46 (-1.7)
	—	0.913 ( 78.69)	0.655 ( 1.65)	1.61	0.89	0.92 (3.4)	-0.49 (-1.9)

Note:  $t$ -values are in parentheses.

the amount of capital gains in year  $t$ . Furthermore, in our specification of the wealth approach, consumer wealth and expected capital gains affect consumption equally:  $W(L)G_t$  and  $W_{t-1}$  have the same coefficient in equation (8). Therefore, we estimate (8) subject to this constraint, using the iterative procedure described above for the income approach. In other words, we let  $\alpha$  and  $\lambda$  vary over a prespecified range, calculate  $Y_t^w$  and  $W(L)G_t$ , compute  $W_t^*$  by adding  $W_{t-1}$  and  $W(L)G_t$  as indicated in equation (7), estimate equation (6), and select regression coefficients which minimize the residual sum of squares. (See fn. 17.) Results for equation (6), for both accrued and realized gains, are reported in Table 7.

Coefficients and weights for individual years are presented in Table 8.

The results are much sharper than in the income approach: the Durbin-Watson statistic suggests serial independence of residuals in all cases,  $W_t^*$  and  $Y_t^w$  are highly significant, the coefficients have the expected signs (positive), and their magnitudes are not affected much by whether accrued or realized gains are included in the equation. The normal rate of return on wealth ( $\beta = k\beta/k$ ) is estimated between .06 and .08, which seems quite reasonable. All in all, equation (6) appears to be much better specified than equation (4). Since  $k\beta > 0$  and significant, we accept  $H3$ : both accrued and realized gains affect consump-

TABLE 7—ESTIMATES OF THE COEFFICIENTS OF THE  
CONSUMPTION FUNCTION, EQUATION (6)

	Constant	$Y_t^w$ ( $k$ )	$W_t^*$ ( $k\beta$ )	$SEE$	$R^2$	$D.W.$
Accrued Gains	4.490 (0.50)	0.708 ( 5.25)	0.062 (3.23)	2.08	0.999	1.94
	—	0.783 (19.79)	0.052 (7.26)	2.02	0.999	1.77
Realized Gains	2.685 (0.30)	0.760 ( 5.65)	0.055 (2.89)	2.12	0.999	1.82
	—	0.801 (21.16)	0.050 (7.15)	2.03	0.999	1.81

Note:  $t$ -values are enclosed in parentheses. Effective number of observations in each case is 13.

TABLE 8—COEFFICIENTS FOR INDIVIDUAL YEARS<sup>a</sup>

<i>i</i>	Realized Gains				Accrued Gains			
	Income		Gains		Income		Gains	
	$x_i$	$y_i$	$w_i$	$g_i$	$x_i$	$y_i$	$w_i$	$g_i$
0	0.47	0.360	0.54	0.030	0.48	0.376	0.27	0.014
1	0.32	0.245	0.32	0.018	0.32	0.251	0.25	0.013
2	0.15	0.115	0.12	0.007	0.14	0.110	0.21	0.011
3	0.05	0.038	0.03	0.002	0.04	0.031	0.16	0.008
4	0.01	0.008	0	0	0.01	0.078	0.12	0.006

<sup>a</sup> The coefficients correspond to equation (6), without its constant term, as reported in Table 7.

tion via wealth.  $H_2$ , however, is rejected because the coefficient of gains is always smaller than that of expected income.

#### *Coefficients for Individual Years*

In equation (6),  $k$  and  $k\beta$  are coefficients of expected income and expected gains, respectively. The expected variables, in turn, are estimated as weighted moving averages of past income and gains. Using  $k$ ,  $k\beta$ , and the weights, we can estimate the income and gains coefficients for individual years. Let  $x_i$  represent the weight and  $y_i$  the corresponding coefficient for income in year  $i$ . Similarly, let  $w_i$  and  $g_i$ , respectively, be the weight and coefficient for gains in year  $i$ . Then,

$$Y_t^w = \sum_{i=0}^4 x_i Y_{t-i}$$

$$W(L)G_t = \sum_{i=0}^4 w_i G_{t-i}$$

$$y_i = kx_i$$

and

$$g_i = k\beta w_i$$

Estimates of  $x_i$ ,  $y_i$ ,  $w_i$ , and  $g_i$  are presented in Table 8.

Note in Table 8 that the coefficient of income is always larger than the coefficients of both realized and accrued gains. Gains realized during the past one or two years affect consumption more than gains accruing during the same period. These results are also reflected in the mean lag for

the three variables: about one year for income, about three quarters for realized gains, and more than two and one-half years for accrued gains.<sup>20</sup> The overall effect of realized gains on consumption is smaller than that of income, but it occurs with a shorter average lag.<sup>21</sup>

#### IV. Economic Implications

The findings reported above have important implications for taxation and other areas of economic analysis and policy. At present, capital gains enjoy special tax treatment: gains are taxed on realization, the tax rates are lower than those on other forms of income and accrued gains on assets transferred by bequest escape taxation altogether. This policy is supported, *inter alia*, by the argument that gains do not significantly affect consumption; there-

<sup>20</sup> The mean of a geometrically declining lag distribution

$$Z_i^e = (1 - \lambda) \sum \lambda^i Z_{t-i}$$

is equal to  $\lambda/(1 - \lambda)$ . Compare Griliches, p. 19.

<sup>21</sup> In estimating equations (3) and (6), we have used finite numerical approximations for infinite lags because of the very small sample at our disposal. The weights in Table 8, however, become very small after 3 periods. Moreover, we reestimated these equations by varying the lags—increasing it to 6 periods for gains, and reducing it to 4 periods for income—the overall results did not change much. Therefore, it is reasonable to expect that the numerical approximations used above did not cause any substantive errors in the final results.

fore, the capital gains tax is paid largely out of savings.<sup>22</sup> Our analysis shows that capital gains, both realized and accrued, affect consumption significantly. Besides, in all likelihood, people treat realized gains like any other income. The present taxation of capital gains has been strongly criticized for its inequity and economic effects.<sup>23</sup> The results presented above question one of the strongest economic arguments in favor of continuing this tax policy.

It is often suggested that either gains should be taxed on accrual or realized gains should be taxed more heavily. If gains are taxed on accrual, other things being equal, larger amounts of gains will be realized. An increase in the rate of tax on realized gains, without a tax on accruals, will discourage realization of gains, but accrued gains will tend to increase. The coefficient of realized gains is larger than that of accrued gains in most cases. Therefore, it is likely that these two policies would have different effects on consumption. Moreover, the coefficient of realized gains is not much different from that of expected income in equation (3); thus, an increase in capital gains tax or the personal income tax would reduce consumption by roughly the same magnitude. This has a clear bearing on stabilization policy.<sup>24</sup>

Our formulation of the consumption function is very similar to that used by Ando and Modigliani:

$$(9) \quad C_t = \alpha_1 Y_t + \alpha_2 Y_t^e + \alpha_3 A_{t-1},$$

<sup>22</sup> Compare Henry Wallich, pp. 140-41.

<sup>23</sup> See, for example, Martin David, chs. 1, 5, and 10.

<sup>24</sup> The model as stated, however, cannot be used to compare the relative stabilizing influence of the personal income tax and the capital gains tax. Suppose gains are taxed more heavily; corporate retained earnings, which are the main source of stock market gains, would be reduced, but dividends, and hence disposable income (excluding capital gains) would increase. The effect on consumption therefore, is ambiguous. A fuller treatment of the role of capital gains in stabilization policy will be presented in a forthcoming paper.

where  $C_t$ ,  $Y_t$ ,  $Y_t^e$  and  $A_{t-1}$  represent aggregate consumption, current nonproperty income, expected nonproperty income, and net worth, respectively. They test the hypothesis that expected nonproperty income is the same as actual current income, except for a possible scale factor, i.e.,

$$Y_t^e = \beta' Y_t; \beta' \simeq 1$$

Alternatively, if  $\alpha_1 \simeq \alpha_2$ , we can combine  $Y_t$  and  $Y_t^e$ . Using the former result, equation (9) can be rewritten as

$$(10) \quad C_t = \left( \frac{1}{\beta'} + \alpha_2 \right) Y_t^e + \alpha_3 A_{t-1}$$

But  $A_{t-1}$  is based on the average price during the year. We can, therefore, isolate the capital gains implicit in  $A_{t-1}$  by evaluating net worth at prices prevailing at the beginning of the period:

$$(11) \quad A_{t-1} = W_{t-1} + \gamma' G_t,$$

where  $W_{t-1}$  is net worth computed from asset prices at the beginning of the year, and  $\gamma'$  represents the fraction of current period gains concealed in  $A_{t-1}$ . Substituting equation (11) into (10), we get:

$$(12) \quad C_t = \delta_1 Y_t^e + \delta_2 W_{t-1} + \delta_3 G_t$$

where

$$\delta_1 = \frac{1}{\beta'} + \alpha_2, \quad \delta_2 = \alpha_3, \quad \text{and} \quad \delta_3 = \alpha_3 \gamma'$$

Apart from the restrictions on parameters, and the use of  $G_t$  instead of  $W(L)G_t$ , equation (12) is of the same form as (8). It is well known that a consumption function of this type can explain both the long-run stability and the cyclical variability of the saving-income ratio observed in empirical studies of the U.S. economy.<sup>25</sup> However, as I suggested in my 1970 paper, if capital gains are included in personal income,

<sup>25</sup> Compare Ando and Modigliani, pp. 76-79 and the references cited there.

they will have to be included in personal saving also. The resulting saving-income ratio shows much greater cyclical variation than that recorded in national income statistics. The main purpose of this paper has been to analyze the effect of capital gains on aggregate consumption; the behavior of the saving-income ratio per se does not concern us here, but the results derived above would be useful in explaining changes in the saving-income ratio.

## REFERENCES

- A. Ando and F. Modigliani, "The 'Life-Cycle' Hypothesis of Saving: Aggregate Implications and Tests," *Amer. Econ. Rev.*, Mar. 1963, 53, 55-84.
- A. Ando and E. C. Brown, "Lags in Fiscal Policy," in E. C. Brown et al., eds., *Stabilization Policies: Commission on Money and Credit*, Englewood Cliffs, 1963.
- J. J. Arena, "Capital Gains and the 'Life-Cycle' Hypothesis of Saving," *Amer. Econ. Rev.*, Mar. 1964, 54, 107-11.
- , "Postwar Stock Market Changes and Consumer Spending," *Rev. Econ. Statist.*, Nov. 1965, 47, 379-91.
- , "The Wealth Effect and Consumption, A Statistical Enquiry," *Yale Econ. Essays*, fall 1963, 3, 251-303.
- R. J. Ball and P. S. Drake, "The Relationship between Aggregate Consumption and Wealth," *Int. Econ. Rev.*, Jan. 1964, 5, 63-81.
- K. B. Bhatia, "Individuals' Capital Gains in the United States: An Empirical Study, 1947-1964," unpublished doctoral dissertation, Univ. Chicago 1969.
- , "Accrued Capital Gains, Personal Income and Saving in the United States, 1948-64," *Rev. Income and Wealth*, Dec. 1970, 16, 363-78.
- , "Capital Gains and the Distribution of Personal Income," mimeo., Jan. 1972.
- R. Bodkin, "Windfall Income and Consumption," *Amer. Econ. Rev.*, Sept. 1959, 49, 602-14.
- D. Cochrane and G. H. Orcutt, "Applications of Least Squares Regression to Relationships Containing Autocorrelated Error Terms," *J. Amer. Statist. Ass.*, Mar. 1949, 44, 32-61.
- G. D. Craig, "Predictive Accuracy of Aggregate Quarterly and Annual Saving Functions," unpublished doctoral dissertation, Univ. Illinois, Urbana 1968.
- M. David, *Alternative Approaches to Capital Gains Taxation*, Washington 1968.
- M. Friedman, *A Theory of the Consumption Function*, Princeton 1957.
- Z. Griliches, "Distributed Lags: A Survey," *Econometrica*, Jan. 1967, 35, 16-49.
- R. M. Haig, "The Concept of Income—Economic and Legal Aspects," in R. Musgrave and C. Shoup, eds., *AEA Readings in the Economics of Taxation*, Homewood 1959.
- J. R. Hicks, *Value and Capital*, Oxford 1946.
- J. Johnston, *Econometric Methods*, New York 1963.
- L. Klein and N. Liviatan, "The Significance of Income Variability on Savings Behavior," *Bull. Oxford Inst. Econ. Statist.*, May 1957, 19, 151-60.
- L. M. Koyck, *Distributed Lags and Investment Analysis*, North-Holland 1954.
- M. Landsberger, "The Life-Cycle Hypothesis: A Reinterpretation and Empirical Test," *Amer. Econ. Rev.*, Mar. 1970, 60, 175-83.
- N. Liviatan, "Consistent Estimation of Distributed Lags," *Int. Econ. Rev.*, Jan. 1963, 4, 44-52.
- E. Malinvaud, *Statistical Methods of Econometrics*, North Holland 1970.
- M. B. McElroy, "Capital Gains and the Theory and Measurement of Income," unpublished doctoral dissertation, Northwestern Univ. 1970.
- J. Mincer, "Employment and Consumption," *Rev. Econ. Statist.*, Feb. 1960, 42, 20-26.
- F. Modigliani and R. Brumberg, "Utility Analysis and the Consumption Function: An Interpretation of Cross Section Data," in K. Kurihara, ed., *Post-Keynesian Economics*, New Brunswick 1954.
- M. Nerlove, *Distributed Lags and Demand Analysis*, Agriculture Handbook No. 141, U.S.D.A., Washington 1958.
- H. C. Simons, *Personal Income Taxation*, Chicago 1938.

- A. Spiro, "Wealth and the Consumption Function," *J. Polit. Econ.*, Aug. 1962, 70, 339-54.
- H. C. Wallich, "Taxation of Capital Gains in the Light of Recent Economic Developments," *Nat. Tax J.*, June 1965, 18, 133-50.
- C. Wright, "Estimating Permanent Income: A Note," *J. Polit. Econ.*, Oct. 1969, 73, 845-50.
- A. Zellner, D. S. Huang, and L. C. Chau, "Further Analysis of the Short-Run Consumption Function, with Emphasis on the Role of Liquid Assets," *Econometrica*, July 1965, 33, 571-81.
- Council of Economic Advisers, *Economic Report of the President*, Washington 1969.
- U.S. Internal Revenue Service, *The Statistics of Income, Individual Income Tax Returns*, various years.

# Keynes-Wicksell and Neoclassical Models of Money and Growth

By STANLEY FISCHER\*

The essential features of Keynes-Wicksell (henceforth KW) monetary growth models, distinguishing them from neoclassical models, are the specification of an independent investment function and the assumption that prices change only in response to excess demand in the goods market.<sup>1</sup> In neoclassical monetary growth models, by contrast, there is no independent investment function and all markets are continuously in equilibrium.

In KW models a steady state of inflation requires persistent excess demand in the goods markets. This suggests that the steady-state properties of such models are unsatisfactory. In neoclassical models, an instantaneous doubling of the quantity of money, however the money is distributed, produces an instantaneous doubling of the price level so long as the expected growth rate of the money supply is the same before and after the "blip" in the money supply. This—according to KW theorists—suggests that there is something amiss in the short-run dynamics of the price level in such models.

In this paper, the price dynamics of both models are discussed, and a modified price determination equation is incorporated

into a KW model. The standard comparative dynamic exercises for monetary growth models are undertaken in this modified model; the modification of the price adjustment equation ensures steady state equilibria rather than disequilibria. The properties of the modified KW model are then compared with those of neoclassical models. Essentially, familiar short-run macro-economic conclusions emerge from consideration of short-run behavior in the modified model and neoclassical conclusions emerge from analysis of its long-run behavior.

## I. Price Dynamics

KW models use the Law of Supply and Demand to determine the rate of inflation.<sup>2</sup> Specifically, it is assumed in KW models that

$$(1) \quad \pi = \lambda(D - S), \quad 0 < \lambda < \infty$$

where  $\pi$  is the rate of inflation,  $D$  and  $S$  are aggregate demand for and supply of goods, each in real terms, and  $\lambda$  is a constant. It is apparent that there cannot be inflation without excess demand if equation (1) determines the rate of inflation, and thus a steady state with inflation requires persistent excess demand. KW models can accordingly have steady states in which individuals are continually frustrated in

\* Assistant professor, department of economics, University of Chicago. I would like to thank George Borts, Rudiger Dornbusch, and Jerome Stein for their helpful comments on an earlier draft. Thanks for comments and discussion are due, too, to William Brock, Jacob Frenkel, Merton Miller, Michael Mussa, Douglas Purvis, and Richard Zecher.

<sup>1</sup> Jerome Stein—who is apparently responsible for the KW designation—has recently provided two very useful expositions of these models (1969, 1970). An earlier article of his (1966), using a KW model which is not so named, provides a full dynamic analysis for the typical KW model.

<sup>2</sup> See Kenneth Arrow. It will be assumed that the reader is familiar with both types of monetary growth models. A two-asset (money and capital), one-sector model is used as the paradigm of neoclassical models (see James Tobin and Miguel Sidrauski); places where my conclusions would differ if some other neoclassical model were used are footnoted. My paradigmatic KW model is contained in Stein's 1969 article.

obtaining the goods they demand, even though their demands are based on correct expectations and perceptions of the price level—and they are condemned to be so frustrated forever after. This is an unappealing result and there are two possible lines of attack on the problem: first, demands could be expected to change in response to such frustrations; alternatively, the price determination equation might be inadequate. I pursue the second approach.

The question raised by (1) and similar equations is: Whose behavior do such equations describe? The standard Walrasian answer is “the auctioneer”; another frequent answer is “somewhat less than competitive firms.”

Consider the auctioneer explanation first. In the standard single period exchange model, the auctioneer calls out prices for each good sequentially on the basis of the mechanism:

$$(2) \quad p_{i,j} = p_{i-1,j} + x_j(p_{i-1})$$

where  $i$  is the iteration number of the current call,  $j$  is the number of the good,  $p$  is the vector of prices, and  $x_j(p_{i-1})$  is an increasing function of excess demand for good  $j$  at the previously called price vector. In intertemporal models an equilibrium price vector is obtained by the above process at the beginning and no further tatonnement is required. If new information is available in each period, as in models including uncertainty, one supposes that there is an “auction” each period. The goal of the auctioneer in each period is to establish market-clearing prices—prices at which demands are equal to supplies.

Equation (1) is an attempt to use (2) in a temporal context so that the  $i$  subscript becomes a  $t$ , and to apply (2) to the aggregate price level. But it ignores the motive of the auctioneer. If the auctioneer expects the general price level at time  $t$  to be different from that at  $t-1$ , then he might use

as his rule of thumb

$$(3) \quad p_{t,j} = p_{t-1,j} \left( \frac{\bar{p}_t}{\bar{p}_{t-1}} \right) + x_j(p_{t-1})$$

where  $\bar{p}_t$  is the general price level expected to prevail at  $t$ , and  $\bar{p}_{t-1}$  is the general price level at  $t-1$ . Aggregating over goods, and in continuous time, an analogue of (3) is:

$$(4) \quad \pi = \pi^* + \lambda(D - S)$$

where  $\pi$  is the actual rate of inflation, and  $\pi^*$  is the expected rate of inflation.

The auctioneer is not present in most markets and it is somewhat unsatisfactory to discuss reasonable behavior for a non-existent economic agent. Consider alternatively the explanation in terms of the behavior of price-setting firms. As suggested by Arrow, and developed by Robert Barro in a recent and interesting paper, since the existence of disequilibrium is inconsistent with certain assumptions of the perfectly competitive model,<sup>3</sup> we may expect price-setting by firms even in industries for which the competitive model is adequate for comparative static analysis.

Barro analyzes optimal price-setting behavior for a monopolistic firm faced with uncertain demand and a fixed cost of adjusting its selling price; the optimal policy is to adjust price only when excess demand or supply reaches certain barriers. He then shows that, by aggregating over firms, the average price may be expected to behave according to (1). Barro confines himself to cases where the aggregate price level is expected to remain constant. Suppose now that all prices but the monopolist's price were expected to increase at the rate  $\pi^*$ ; then costs would be expected to rise at the rate  $\pi^*$  (since the cost function

<sup>3</sup> In particular, in disequilibrium it cannot be true that each firm can sell as much as it wants at the going price and each consumer can purchase as much as he wants at the going price.

is homogeneous of degree one in prices), and as of any given price fixed over an interval by the monopolist, the relative price of the monopolist's output would be falling at the rate  $\pi^*$ . Then, in adjusting prices, the monopolist could be expected to include an adjustment for the trend in prices over the period for which he expects to keep his own price constant. Aggregating over firms, one would expect to reach an equation similar to (4).

Thus, on either score, an equation such as (4) is a more adequate representation of price adjustment than is (1). Accordingly, I proceed in Section II to an analysis of a KW model incorporating equation (4). Stein (1970) has in fact suggested that an equation like (4) might be useful in reconciling KW and neoclassical models. Similar equations may be found to describe wage and price adjustment in the literature.<sup>4</sup>

Before presenting the modified KW model, it is necessary to discuss the price dynamics implicit in the usual neoclassical model. The per capita demand for real balances ( $m^d$ ) is a function of the per capita capital stock ( $k$ ) and the expected rate of inflation ( $\pi^*$ ):

$$(5) \quad m^d = L(k, \pi^*), \quad L_1 > 0, L_2 < 0$$

At any instant of time the capital stock (we omit "per capita" where no confusion is likely to result) and the expected rate of inflation are given, as is the nominal money stock and population. Then, adding to (5) the neoclassical specification that the money market is always in equilibrium

$$(6) \quad M/PN \equiv m = m^d$$

is sufficient to determine the price level. In particular, a doubling of the stock of money will double the price level but leave the system otherwise unaffected.<sup>5</sup>

Is there any reason to regard this instantaneous neutrality with suspicion? There are circumstances under which it might be regarded as reasonable: for instance, if it was announced that at some point of time every individual's nominal money balances would be doubled, then, given some sophistication by economic agents, it might be realized that this action was analogous to creation of a new unit of account and the price level might simply double. It is, however, a basic assumption of neoclassical models that injections of money are not distributed on the basis of existing holdings of money (since otherwise the transfer payments by which the money supply is expanded would be equivalent to interest payments on money holdings). Given this assumption, increases in the nominal balances of some individuals in the economy can be expected to produce their effects on prices gradually, through real balance effects, rather than instantaneously. Hence the KW objection to this neutrality has force.

Using (5) and (6), the rate of inflation in neoclassical models is given by

$$(7) \quad \pi = \mu - n - \frac{1}{m} [L_1 Dk + L_2 D\pi^*]$$

where  $\mu$  is the (assumed constant) rate of expansion of the nominal money supply,  $n$  is the rate of population growth, and  $D$  denotes the time derivative. In the steady state  $\pi = \mu - n$ ; thus the rate of inflation will be reduced below its steady-state value by capital accumulation and raised above its steady-state value by increases in the expected rate of inflation. Even leaving aside the expectational factor,  $D\pi^*$ , equation (7) is not analogous to (4).

<sup>4</sup> See, for example, Edmund Phelps.

<sup>5</sup> In two-sector neoclassical models (e.g., Duncan Foley and Miguel Sidrauski) determination of the price level requires also commodity market clearing, and the

price level cannot be said to be determined by the requirement of portfolio balance. It remains true that in such models, "jumps" in the money stock affect only the aggregate price level.

## II. The Modified KW Model

In outlining this KW model I shall point to its departures from neoclassical analysis. Both types of model have in common a production function, stock demand functions for assets, a savings function, and an expectations function. I shall specify forms of these functions which could be used in either type of model.

The per capita output of goods is

$$(8) \quad y = f(k) \quad f' > 0, f'' < 0$$

where, for convenience, it is assumed the Inada conditions hold and that real balances do not enter the production function. It is also assumed that the labor force, growing at the rate  $n$ , is supplied inelastically and that full employment is maintained.<sup>6</sup>

There are three assets: money, private bonds, and physical capital. Stock demand functions for real balances, real bonds (the excess demand function, since it is assumed there are no outside bonds), and capital are given by (9), (10), and (11), respectively.<sup>7</sup> The assets are assumed to be gross substitutes. The variable  $y$ , output, enters to represent the transactions demand for money. Per capita wealth,  $a = (k + m)$ ,

<sup>6</sup> For a KW model with variable employment, see Keizo Nagatani.

<sup>7</sup> The demand functions for assets differ from those used in Foley and Sidrauski only in that the price of capital does not enter. It is assumed that production always takes place away from corners of the production possibility frontier so that the relative price of capital and consumption goods remains fixed. I note, quoting David Levhari and Don Patinkin, "that it would be more consistent with general considerations of economic theory if . . . [the demands for assets] . . . were represented as depending upon disposable income . . . This, however, would greatly complicate the . . . analysis which follows. . . ." (p. 720).

enters as the stock budget constraint.<sup>8</sup> Bonds and capital are not perfect substitutes so that  $f'(k) + \pi^*$ , the expected nominal return on capital, may differ from  $\rho$ , the nominal interest rate. The three demand functions are dependent since the sum of the demands for assets is constrained by wealth at each instant.

Per capita savings is a function of disposable income and wealth:

$$(12) \quad s = s(y^e, a), \quad 1 > s_1 > 0, s_2 < 0$$

Expected disposable income,  $y^e$  consists of factor payments,  $f(k)$ , plus transfer payments  $\mu m$ , where  $\mu$  is the constant and preannounced rate of expansion of the nominal money supply (it is assumed that the current price level is correctly perceived), minus expected capital losses on money holdings,  $\pi^* m$ . Thus

$$y^e = f(k) + (\mu - \pi^*)m$$

Saving is definitionally equal to desired additions to asset holdings; it is the sum of  $l^d$ ,  $h^d$ , and  $x^d$  which are desired additions, per capita, to real balances, bonds and capital, respectively. Consumption demand and savings demand are constrained by disposable income:

$$(13) \quad y^e = c^d + s$$

It is well known that the stability of dynamic models is heavily dependent on the expectations function. We assume here adaptive expectations:

$$(14) \quad \pi^* = \beta(\pi - \pi^*), \quad 0 < \beta < \infty$$

Thus far we have outlined a fairly stan-

<sup>8</sup> Since there are no outside bonds in the model, the net per capita value of bonds is zero.

$$(9) \quad m^d = L(y, a, f'(k) + \pi^*, \rho) \quad L_1 > 0, 1 > L_2 > 0, L_3 < 0, L_4 < 0$$

$$(10) \quad b^d = H(y, a, f'(k) + \pi^*, \rho) \quad H_1 < 0, H_2 > 0, H_3 < 0, H_4 > 0$$

$$(11) \quad k^d = J(y, a, f'(k) + \pi^*, \rho) \quad J_1 < 0, 1 > J_2 > 0, J_3 > 0, J_4 < 0$$

dard neoclassical model. A neoclassical analysis would proceed as follows: assume asset market equilibrium and use any two of (9)–(11) to determine the price level and the nominal interest rate at each instant of time—these are functions of the capital stock and expected rate of inflation. Then assume that consumption demand is always satisfied and obtain the rate of capital accumulation as the residual of output minus consumption.

The scene is then set for determining “next instant’s” short-run equilibrium; the economy proceeds through these equilibria, and if it is stable, ultimately reaches a steady state in which the capital stock and expected rate of inflation are constant. In fact, the model we have set up is very similar to Levhari and Patinkin’s “Money as a Consumer Good” model.

The four KW features of the model follow. First, there is the specification of an investment demand function,  $x^d$ . We assume a stock adjustment demand for investment.

$$(15) \quad x^d = nk + \Phi(k^d - k), \quad \Phi' > 0$$

The flow demand for capital consists of the replacement demand,  $nk$ , plus a term which depends on the divergence between the actual capital stock and that demanded at the current levels of wealth and current rates of return and income. The basic justification for (15) lies in the existence of adjustment costs in changing the capital stock: the greater the divergence between actual and desired capital stocks, the greater the costs that can profitably be incurred in changing the capital stock.<sup>9</sup>

The investment demand function (15) has the property—which is the basis for

investment functions in Stein’s KW models—that an increase in the difference between the expected nominal return on capital,  $f'(k) + \pi^*$ , and the nominal interest rate,  $\rho$ , increases investment demand. This is the “Wicksell” feature of KW models for

$$(16) \quad f'(k) + \pi^* - \rho = f'(k) - (\rho - \pi^*);$$

the first term on the right-hand side of (16) is the natural rate and the second is the real rate, and differences between these two rates affect investment demand.<sup>10</sup>

Second, there is the price adjustment equation, in which it remains to specify aggregate demand and supply.

$$(17) \quad \pi = \pi^* + \lambda(c^d + x^d - f(k))$$

The demand for goods consists of the demands for consumption and investment; the supply is simply full employment output.

Third, it is specified that the bond market be continuously in equilibrium, so that

$$(18) \quad b = b^d = 0$$

This is an assumption of convenience rather than necessity.<sup>11</sup>

Fourth, there is the question of the allocation of output in periods of excess demand or supply. Here it is assumed that both consumption and investment plans are partially frustrated when there is excess demand; in particular, planned investment is reduced by some positive fraction  $(1 - \gamma)$  of excess demand to give the actual rate of investment.

$$(19) \quad x = x^d - (1 - \gamma)[c^d + x^d - f(k)], \\ 0 < \gamma < 1$$

In general  $\gamma$  could be expected to be an endogenous variable rather than a con-

<sup>9</sup> See Robert Eisner and Robert Strotz for the derivation of an investment demand function such as (15); see also Marc Nerlove for critical comments on this and subsequent developments. Note that although adjustment costs are invoked in explaining (15), they are not explicitly incorporated in the model.

<sup>10</sup> The “Keynes” part lies in the specification of an independent investment demand function; other Keynesian features, such as unemployment, can be captured in KW models with variable employment.

<sup>11</sup> In Stein (1966), for instance, it is assumed that the money market is in equilibrium.

stant; while (19) is very much a *deus ex machina*, theories of allocation under disequilibrium are not well developed and there is no formulation which is obviously theoretically superior at this stage. Note that (19) is equivalent, through (17), to

$$(20) \quad x = x^a - \frac{1 - \gamma}{\lambda} (\pi - \pi^*)$$

Before proceeding to an exposition of the short- and long-run properties of the modified model, we use the assumption that the bond market is always in equilibrium (18), to derive the implied relationship between the nominal rate and the capital stock, real balances, and the expected rate of inflation. Given  $k$ ,  $m$ , and  $\pi^*$ , there is, from (10) and (18), only one nominal interest rate which equilibrates the bond market. Specifically

$$(21) \quad \rho = A(k, m, \pi^*)$$

where

$$A_1 = \frac{-1}{H_4} [H_1 f' + H_2 + H_3 f''] \stackrel{?}{<} 0$$

$$A_2 = \frac{-H_2}{H_4} < 0$$

$$A_3 = \frac{-H_3}{H_4} > 0$$

The only ambiguity in (21) concerns the effects of an increase in the capital stock on the nominal rate: there is, in addition to the substitution effect ( $H_3 f''$ ) and wealth effect ( $H_2$ ), an income effect, ( $H_1 f'$ ); we assume that the substitution and wealth effects dominate and that the reduced real rental on capital resulting from an increase in  $k$  leads to a decrease in the nominal rate as of any given  $\pi^*$ . Thus, we assume that increases in the capital stock tend to reduce the nominal rate; our earlier assumptions imply that increases in real balances tend to reduce the nominal rate

while increases in the expected rate of inflation tend to increase the nominal rate of interest.

### III. The Short and Long Run in the Modified Model

We now discuss the behavior of this KW model in the short and long run. Given the assumption that the adjustment coefficient,  $\lambda$ , in (17) is finite, the price level is given at any instant—that is, it is inherited from the past. Accordingly,  $m$ , real balances per capita, is determined exogenously, for  $M$ , nominal balances, is a policy variable. The capital stock and the expected rate of inflation are also inherited from the past. Thus, at an instant of time,  $k$ ,  $m$ , and  $\pi^*$  are predetermined.

The behavioral relations of the model determine, in the short run, the nominal rate of interest and thence, through the goods market, the rate of inflation. Given the rate of inflation, and  $k$ ,  $m$ , and  $\pi^*$ , the rate of capital accumulation is determined from (19), and the rate of change of the expected rate of inflation from (14). The stage is then set to determine the capital stock, real balances, and the expected rate of inflation at the next “instant”; the economy proceeds in this way through time, reaching a steady state if the system is stable. The remainder of this section consists of a more detailed examination of this process.<sup>12</sup>

Given  $k$ ,  $m$ , and  $\pi^*$ , the predetermined variables, the nominal interest rate is determined through the requirement of bond market equilibrium, and is given by (21). That nominal rate in turn, together with the predetermined variables, determines the demands for consumption and investment and the consequent rate of inflation.

<sup>12</sup> The verbal description we give of the dynamic process of this economy corresponds more closely to a difference equation system than to the differential equation system contained in the formal analysis; this is simply a matter of convenience.

Using (17) and the flow budget constraint, the rate of inflation is

$$(17') \quad \pi = \pi^* + \lambda(x^d + (\mu - \pi^*)m - s(y^e, a))$$

Consider now the effects of changes in  $k$ ,  $m$ , and  $\pi^*$  on the rate of inflation. The effects of changes in  $k$  and  $m$  occur only insofar as excess demand is affected (recall that  $x^d$  is a function of the nominal rate, so that effects working through the bond market must also be considered) while a change in  $\pi^*$  has an expectational effect on the rate of inflation in addition to excess demand effects. We obtain

$$(22) \quad \pi = G(k, m, \pi^*, \mu)$$

where

$$G_1 = \lambda \left( \Phi' \left( \frac{dJ}{dk} - 1 \right) + n - s_1 f' - s_2 \right)$$

$$G_2 = \lambda \left( \Phi' \frac{dJ}{dm} + (\mu - \pi^*)(1 - s_1) - s_2 \right) > 0$$

$$G_3 = 1 + \lambda \left( \Phi' \frac{dJ}{d\pi^*} - m(1 - s_1) \right) ? > 0$$

$$G_4 = \lambda m(1 - s_1) > 0$$

The derivatives of the  $J$  function are written as total derivatives to indicate that bond market effects are to be included.

Increases in the capital stock have an uncertain effect on excess demand; they reduce the stock excess demand for capital<sup>13</sup> but may either increase or decrease savings since the income and wealth effects on savings work in opposite directions. If the system is near the golden rule, then  $n - s_1 f' > 0$  and the term  $(n - s_1 f' - s_2)$  will be positive. Thus the sign of  $G_1$  is ambiguous.

Increases in real balances are inflation-

ary; they increase both consumption and investment demand. Increases in the expected rate of inflation have a direct effect on actual inflation throughout the expectations effect—they also increase investment demand but reduce consumption demand by reducing the value of expected transfer payments. Thus, whether the actual rate of inflation increases by more or less than the expected rate depends on whether increases in the expected rate produce an excess supply or excess demand for goods; in other words, on whether the reduction in consumption demand is greater than or less than the increase in investment demand. It later turns out that this is an important factor in determining the stability of the system, and it may be seen that the smaller is  $\Phi'$ —the more slowly is the capital stock adjusted—the more likely is  $(G_3 - 1)$  to be negative. Finally, an increase in the rate of growth of the money stock increases transfer payments and is inflationary.

The "short-run" position of the economy is determined by (21) and (22). Its behavior through time is determined by the capital accumulation equation (20), the rate of change of real balances equation which can be derived by differentiating  $m$  with respect to time, and the expectations equation (14). For convenience we rewrite and renumber these equations here:

$$(23) \quad Dk = \Phi[J(y, a, f'(k) + \pi^*, A(\cdot)) - k] - \frac{1 - \gamma}{\lambda} [G(k, m, \pi^*, \mu) - \pi^*]$$

$$(24) \quad Dm = [\mu - n - G(k, m, \pi^*, \mu)]m$$

$$(25) \quad D\pi^* = \beta[G(k, m, \pi^*, \mu) - \pi^*]$$

Consider now the steady state for this economy. In the steady state,  $D\pi^* = 0$ , and so the actual rate of inflation is equal to the expected rate, and, from (24), each is equal to  $(\mu - n)$ . From (23), the demand for the

<sup>13</sup> This may be shown by computing the derivative  $dJ/dk - 1$  and using the stock budget constraint.

capital stock is equal to the existing capital stock, and there is no excess demand for capital; from the stock budget constraint it follows that there is no excess demand for real balances either. From (17), the excess demand for goods is also zero and since investment demand is satisfied, so is consumption demand.

As in the neoclassical model, there are no unsatisfied demands in the steady state of the modified KW model. The reformulation of the price adjustment equation is thus sufficient to remove the unsatisfactory feature of previously published KW models—the persistence of excess demand in the steady state.

#### IV. Changes in the Stock of Money and in the Rate of Growth of Money Stock

Suppose the economy is in the steady state and there is an increase in the money stock, but no change in the rate of growth of the money supply. Then since  $\mu$  is the only exogenous variable in the system in the long run, it is apparent that if the system is stable, it will return to the same steady state. However, this economy, unlike our earlier neoclassical system, will be forced out of equilibrium by the increase in the money stock, and will take time to return to its steady state. The steady-state neutrality is of course neoclassical—but the dynamics is not.

Consider now the impact effects of an increase in the money stock. The nominal interest rate is reduced, and the rate of inflation is increased because excess demand is increased. The increase in the rate of inflation increases the expected rate of inflation and begins to reduce real balances. The effects of the increase in the money stock on capital accumulation are ambiguous: the demand for both investment goods and consumption goods is increased, and investment is more likely to increase the relatively greater are real balance effects on investment demand and

the more fully are investment plans, rather than consumption plans, realized. This short-run story is very Keynesian insofar as the effects of the change in the money stock manifest themselves in the bond market and result in an increase in investment demand through the lowering of the nominal rate. If we had been dealing with a model with unemployed resources, the story would have been even more Keynesian for the increase in both consumption and investment demand could have called forth more output, rather than resulting in inflation.

The path followed by the economy thereafter depends on its stability properties, which are analyzed in the Appendix. It is shown in the Appendix that if the steady state is near the golden rule capital stock, then a necessary condition for stability is that increases in the expected rate of inflation reduce excess demand—this, as discussed above, is helped by the slow adjustment of investment demand to changes in the desired capital stock, and damaged by a great sensitivity of the demand for capital to the expected rate of inflation. It is also shown that slow adjustment of expectations—as in the neoclassical model—and rapid adjustment of prices to eliminate excess demands are conducive to stability. However, the conditions for the rapid adjustment of expectations—a large  $\beta$ —to produce instability are less stringent than they are in neoclassical models.

Finally, we consider the comparative steady-state properties of the modified KW model. An increase in the growth rate of the nominal money supply ultimately increases the expected rate of inflation by the same amount as the increase in the monetary growth rate. The higher expected rate of inflation increases the demand for capital and reduces the demand for real balances; one of the factors determining the new steady state is thus the asset demand

functions and the fact that there will be no excess demands in the long run; the other factor determining the new steady state is savings behavior. Working with our full system of differential equations, we obtain

$$(26) \quad \frac{dk^*}{d\mu} = -\frac{\beta m \Phi' \lambda}{Z_3} \frac{dJ}{d\pi^*} (n(1-s_1) - s_2)$$

and

$$(27) \quad \frac{dm^*}{d\mu} = \frac{\beta m \Phi' \lambda}{Z_3} \frac{dJ}{d\pi^*} (n - s_1 f' - s_2)$$

where  $Z_3$  is the determinant of the matrix in the Appendix which has to be negative for stability. This negativity is assured if  $(n - s_1 f' - s_2) > 0$ .

Thus we can say that if the system is in a stable steady state, increases in the rate of growth of money unambiguously increase the equilibrium capital intensity; and if that steady state is near the golden rule capital stock, increases in the rate of growth of money reduce equilibrium real balances. In any event, if increases in the capital stock reduce savings, so  $s_1 f' - s_2 < 0$ , then increases in  $\mu$  increase  $k^*$  and reduce  $m^*$ .

These results are familiar and early comparative steady state neoclassical propositions. We obtain them, of course, because this KW system has the same steady-state properties as our neoclassical model of Section II, which was set up to be very similar to earlier neoclassical monetary growth models.<sup>14</sup> Although we chose to represent our steady state by using (23)–(25) we could equally well have been neoclassical and described the steady state in terms of asset market equilibrium and the requirement that savings be just sufficient to maintain real per capita assets constant.

<sup>14</sup> In particular, our use of output rather than disposable income in the asset demand functions, and the omission of imputed interest on real balances enable us to avoid several pitfalls.

It is, incidentally, interesting to use (20) to examine the impact effect on investment of an increase in the growth rate of the money supply. The demand for investment goods  $x^d$  is unaffected by increases in  $\mu$ . Thus the impact effect of a change in  $\mu$  depends only on its effect on the rate of inflation. The rate of inflation increases with  $\mu$ , so that the actual rate of investment falls when  $\mu$  is increased. The increase in  $\mu$  increases consumption demand but not investment demand and so some investment is displaced. Thus, initially the capital stock falls when the rate of growth of money is increased, though ultimately the capital stock increases. This is similar to the behavior of the capital stock following an increase in  $\mu$  in Sidrauski.

## V. Conclusions

The purpose of this paper has been to modify a KW model in a way which removes the feature of steady-state excess demand in such models and to compare the resulting model with a neoclassical model based on the same demand functions for assets and savings. The paper has made it clear that the element producing the unsatisfactory features of KW models is the price adjustment equation, and arguments have been presented for using an alternative adjustment equation in which prices may change because they are expected to change, as well as because there is excess demand. The potential of KW models for a useful theory of short-run dynamics, emphasized by others, has been demonstrated in the context of the modified model. It has also been shown that there is no inherent reason for the long-run properties of KW and neoclassical models to differ, so long as the KW investment demand function is consistent with the neoclassical stock demand function for capital.

$$(A1) \quad Z = \begin{bmatrix} \Phi' \left( \frac{dJ}{dk} - 1 \right) - \frac{1-\gamma}{\lambda} G_1 & \Phi' \frac{dJ}{dm} - \frac{1-\gamma}{\lambda} G_2 & \Phi' \frac{dJ}{d\pi^*} - \frac{1-\gamma}{\lambda} (G_3 - 1) \\ -G_1 m & -G_2 m & -G_3 m \\ \beta G_1 & \beta G_2 & \beta (G_3 - 1) \end{bmatrix}$$

## APPENDIX

*Stability Conditions*

The matrix involved in determining the local stability of the system (23)–(25) is shown in (A1) above.

Let  $Z_1$  be the trace of  $Z$ ,  $Z_2$  the sum of its second-order principal minors, and  $Z_3$  its determinant. Necessary and sufficient conditions for local stability are

$$(A2) \quad \begin{aligned} Z_1 &< 0 \\ Z_3 &< 0 \\ Z_1 Z_2 - Z_3 &< 0 \end{aligned}$$

A necessary condition implied by (A2) is that  $Z_2$  be positive.

Now,

$$(A3) \quad Z_1 = -\beta m (G_3 - 1) G_2 \left[ \Phi' \left( \frac{dJ}{dk} - 1 \right) - \frac{1-\gamma}{\lambda} G_1 \right] < 0$$

From the derivatives given in (22), we know that  $G_2$  is positive; it follows that the product of

$(G_3 - 1)$  and

$$\left[ \Phi' \left( \frac{dJ}{dk} - 1 \right) - \frac{1-\gamma}{\lambda} G_1 \right]$$

must be positive. These are, respectively, the terms  $\partial(D\pi^*)/\partial\pi^*$  and  $\partial(Dk)/\partial k$ . Consider first  $\partial(Dk)/\partial k$  which is

$$(A4) \quad \frac{\partial(Dk)}{\partial k} = \gamma \Phi' \left( \frac{dJ}{dk} - 1 \right) - (1-\gamma)(n - s_1 f' - s_2)$$

The first term in parentheses is negative by

virtue of the gross substitute assumption, and if  $(n - s_1 f' - s_2) > 0$ , the whole expression will be negative. Now, at low levels of the capital stock,  $f'$  is very large and the above expression may be negative unless  $\gamma$  is close to unity; for higher levels of the capital stock, and certainly when it is near the golden rule, we are assured that  $\partial(Dk)/\partial k$  is negative. We shall assume that the steady state about which we are examining the dynamics is such that  $n - s_1 f' - s_2 > 0$  and hence  $\partial(Dk)/\partial k < 0$ .

Given this, it is necessary that  $(G_3 - 1)$  be negative, or that the direct effects of an increase in the expected rate of inflation in the goods market be negative—this requires that the adjustment coefficient in the investment equation,  $\Phi'$ , be sufficiently small and/or that  $dJ/d\pi^*$  be small.

Second

$$(A5) \quad Z_3 = -\beta \Phi' m \left[ G_1 \frac{dJ}{dm} - G_2 \left( \frac{dJ}{dk} - 1 \right) \right] < 0$$

$$= -\beta \lambda \Phi' m \left[ \frac{dJ}{dm} (n - s_1 f' - s_2) - \left( \frac{dJ}{dk} - 1 \right) (n(1 - s_1) - s_2) \right]$$

Given the assumption  $n - s_1 f' - s_2 > 0$ , this is negative.

The value of  $Z_2$  is

$$(A6) \quad Z_2 = m \beta G_2 + \beta \Phi' \left[ (G_3 - 1) \left( \frac{dJ}{dk} - 1 \right) - G_1 \frac{dJ}{d\pi^*} \right] - \frac{Z_3}{\beta} > 0$$

The sign of the bracketed term is ambiguous: after substitution the term becomes

$$-\beta\lambda\Phi'\left[\left(\frac{dJ}{dk}-1\right)(1-s_1)m+\frac{dJ}{d\pi^*}(n-s_1f'-s_2)\right]$$

While the first term within the brackets is negative, the second is positive and potentially destabilizing. Note that I have already discussed the size of  $dJ/d\pi^*$ , for if this term is large, there may be trouble with price level stability (see the discussion after equation (22) and above (A5)). Note also that the smaller is  $\beta$ —the more slowly do expectations adapt—the more likely is this stability condition (A6) to be met.

Finally,

$$(A7) \quad \begin{aligned} & Z_1 Z_2 - Z_3 \\ &= Z_1 \left[ m\beta G_2 + \beta\Phi' \left( (G_3 - 1) \left( \frac{dJ}{dk} - 1 \right) - G_1 \frac{dJ}{d\pi^*} \right) \right] - Z_3 \left[ 1 + \frac{Z_1}{\beta} \right] < 0 \end{aligned}$$

Evidently, the larger is  $Z_1$ , in absolute value, the more likely is the system to be stable provided  $Z_2 > 0$ ; the greater (in absolute value) are  $\partial(Dk)/\partial k$ ,  $\partial G/\partial m$  and  $(\partial G/\partial \pi^* - 1)$ , the more likely is the system to be stable. The last two of these derivatives are increasing functions of  $\lambda$ , and thus the faster does the price level adjust in response to excess demand, the more likely is stability.

#### REFERENCES

- K. J. Arrow, "Toward a Theory of Price Adjustment," in M. Abramowitz, ed., *The Allocation of Economic Resources*, Stanford 1959, 44-51.
- R. J. Barro, "A Theory of Monopolistic Price Adjustment," *Rev. Econ. Stud.*, Jan. 1972, 39, 17-26.
- R. Eisner and R. H. Strotz, "Determinants of Business Investment," in D. B. Suits et al., eds., *Impact of Monetary Policy*, Englewood Cliffs 1963, 59-337.
- D. K. Foley and M. Sidrauski, "Portfolio Choice, Investment and Growth," *Amer. Econ. Rev.*, Mar. 1970, 60, 44-63.
- D. Levhari and D. Patinkin, "The Role of Money in a Simple Growth Model," *Amer. Econ. Rev.*, Sept. 1968, 58, 713-53.
- K. Nagatani, "A Monetary Growth Model with Variable Employment," *J. Money, Credit, Banking*, May 1969, 1, 188-206.
- M. Nerlove, "On Lags in Economic Behavior," *Econometrica*, forthcoming.
- E. S. Phelps, "Money Wage Dynamics and Labor Market Equilibrium," in E. S. Phelps, ed., *Microeconomic Foundations of Employment and Inflation Theory*, New York 1970, 124-66.
- M. Sidrauski, "Inflation and Economic Growth," *J. Polit. Econ.*, Dec. 1967, 75, 796-810.
- J. L. Stein, "Money and Capacity Growth," *J. Polit. Econ.*, Oct. 1966, 74, 451-65.
- , "Neoclassical and Keynes-Wicksell Monetary Growth Models," *J. Money, Credit, Banking*, May 1969, 1, 153-71.
- , "Monetary Growth Theory in Perspective," *Amer. Econ. Rev.*, Mar. 1970, 60, 85-106.
- J. Tobin, "Money and Economic Growth," *Econometrica*, Oct. 1965, 33, 671-84.

# The Demand for the Services of Non-Federal Governments

By THOMAS E. BORCHERDING AND ROBERT T. DEACON\*

The empirical literature relating cross-sectional variations of per capita public spending to various economic, political and demographic factors is both lengthy and varied.<sup>1</sup> Differences in expenditures over political units are explained by differences in per capita incomes, urbanization, area, population density, taxable capacity, tax rates, absolute population size, grants-in-aid from higher levels of government, and school-aged population rates. With few exceptions,<sup>2</sup> the models employed in these studies are *ad hoc* constructions with little basis in the theory of choice.<sup>3</sup>

Our aims are to posit a model of public

spending derived from the received theory of collective decision making<sup>4</sup> and to test the significance of certain variables assumed by this simple theory to be important determinants of the levels of state and local government expenditures.

## I. The Model

To accomplish these twin ends of theoretical consistency and statistical estimation, it is necessary to make certain assumptions that are not fully in accord with reality, but which are exceedingly convenient. These assumptions concern three important factors affecting the outcome of public choice: the rules for aggregating voter preferences, the tastes of the choosers, and the opportunity costs to the choosers of the activities undertaken.

Following Downs and Tullock, we assume that in each political unit a government is elected by majority rule. The voting franchise is general and entry into political activity is both brisk and unrestricted. Competition between political entrepreneurs leads to the election of a government that chooses a platform identical to the optimal position of the median voter.<sup>5</sup>

<sup>4</sup> This is the voluntary exchange theory of the nineteenth century continental scholars mentioned in fn. 3, with the majority rule variant as first discussed by Howard Bowen and later developed by Kenneth Arrow, Duncan Black, Anthony Downs, James Buchanan and Gordon Tullock. In 1966 Davis and Haines listed a rather complete bibliography of this literature. Since then Tullock extended the majority rule model from three to  $n$  persons. See fn. 5.

<sup>5</sup> This assumption is implicit in Bowen. Downs shows that when there are two issues and three voters, implicit logrolling may take place, i.e., political entrepreneurs may take the stance of intense minorities. However,

\* Associate professor, department of economics, Virginia Polytechnic Institute and State University, and assistant professor, department of economics, University of California, Santa Barbara. The research was undertaken while we were associated with the department of economics, University of Washington. The Relm Foundation provided partial research support for Borcharding and the Institute for Economic Research, University of Washington, provided assistance to Deacon. We wish especially to acknowledge our debt to Yoram Barzel and Potluri Rao of the University of Washington without associating them with any errors which may remain. Others who were helpful are Roger L. Miller and Richard Parks of the University of Washington, and Charles Goetz and Gordon Tullock of Virginia Polytechnic Institute and State University.

<sup>1</sup> These studies are surveyed through 1964 in Borcharding. The recent literature on state and local variations in the United States has been partially reported by Werner Hirsch.

<sup>2</sup> These are William Birdsall, Otto Davis and George Haines, James Barr, and Robin Barlow.

<sup>3</sup> As far as we can tell, Ronald Coase was the first to criticize the methodology of this empirical research for its *ad hoc* nature. This is rather odd when it is recalled that both the theoretical model of public choice (Wicksell, Lindahl, Mazzola, de Viti de Marco and Pantaleoni) and the empirical work (Wagner) both began around the same time, the turn of the last century. This seems rather a long time to discover so obvious an anomaly.

Citizens are assumed to be informed about the costs and benefits of government spending. The median voter, whose position is known definitively *ex post* and is speculated upon by political entrepreneurs *ex ante*, chooses the level of spending by voting for candidates who offer him the most efficient set of public services and taxes. This, in turn, implies that successful candidates are those who propose platforms that bring the median voter's marginal tax price in line with his marginal benefit. Logrolling or side payments between voters is assumed to be inefficient because of high transactions costs. The utility functions of the median preference holders are assumed to be similar between political units.

As for the organization and supply of the various governmental services bureaucracy is assumed to be no impediment to efficient production; hence, each output level is produced at least cost. Production functions over political units are identical and are taken to be of the Cobb-Douglas constant returns (C-D) variety. Labor and capital are the only factors of production. Within each political unit both factors are available for public purchase at invariant prices, i.e., they are both available in perfectly elastic supply. However, between political units capital is assumed to be perfectly mobile, whereas labor is not. This implies that the rental price per unit of capital is the same over all units, but the wage rate can differ.<sup>6</sup>

Finally, constitutional restrictions permit only nondiscriminatory taxes and expenditures.

---

Tullock, 37-49, has developed an important core theorem in which the median tastes dominate as long as the number of voters is large relative to the number of issues and the preference peaks of the citizenry are evenly or normally distributed over the issue space.

<sup>6</sup> The reason for the choice of this type of production function and the assumptions about the input prices will be made clear below. This formulation was originally suggested to us by Yoram Barzel (1967a).

Before this model is developed in simple mathematical terms, we shall define the relevant variables for the analysis:

$X$  = the physical output in a particular category

$L$  = labor

$K$  = capital

$w$  = the wage rate of  $L$

$r$  = the rental rate of  $K$

$C_x$  = the marginal cost of  $X$

$q$  = the quantity of  $X$  captured by the median voter

$N$  = the number of citizens in a political unit

$\alpha$  = the degree of divisibility in the consumption of  $X$

$s$  = the marginal tax-price of  $q$  to the median voter

$y$  = the income of the median voter

$e$  = per capita expenditure in a particular category

Using the C-D assumption, output is expressed:

$$(1) \quad X = aL^\beta K^{1-\beta}$$

Using the assumption of efficient production (cost minimization) we obtain:

$$(2) \quad w = \frac{\beta C_x X}{L}$$

$$\text{and} \quad r = \frac{(1 - \beta) C_x X}{K}$$

$$(2') \quad L = \frac{\beta C_x X}{w}$$

$$\text{and} \quad K = \frac{(1 - \beta) C_x X}{r}$$

Substituting  $L$  and  $K$  of (2') into (1), we can express output as

$$(3) \quad \begin{aligned} X &= a \left[ \frac{\beta C_x X}{w} \right]^\beta \left[ \frac{(1 - \beta) C_x X}{r} \right]^{(1-\beta)} \\ &= a C_x X \left( \frac{\beta}{w} \right)^\beta \left( \frac{1 - \beta}{r} \right)^{(1-\beta)} \end{aligned}$$

This yields the following expression for marginal cost

$$(3') \quad C_x = \left(\frac{1}{a}\right) \left(\frac{w}{\beta}\right)^\beta \left(\frac{r}{1-\beta}\right)^{(1-\beta)}$$

Recalling that the rental rate on capital is the same over all units, (3') can be re-written:

$$(3'') \quad C_x = a' w^\beta$$

Thus, (3'') describes a unique and horizontal supply function for each unit dependent only on the wage rate in that unit (as well as  $\beta$  and a constant term which are common to all units).

Because of nondiscrimination in expenditure,  $q$ , the amount of  $X$  captured by the median voter depends only on the divisibility of the service flow of  $X$  in consumption. For instance, if  $X$  is a purely private good, then the median citizen receives an aliquot share of  $X$  equal to  $X/N$ ; whereas, if  $X$  is purely public,  $q$  necessarily equals  $X$ . Equation (4) is an arbitrary but convenient device for handling these extremes as well as intermediate cases where the service flows are partially divisible.

$$(4) \quad q = \frac{X}{N^\alpha}$$

Note that  $\alpha$  is unity when the output is purely private and zero when it is purely public; intermediate values imply quasi publicness or quasi privateness in consumption.

Nondiscrimination in taxation requires that the median preference holder pays an equal share of taxes to finance each unit of output,  $X$ , produced, i.e.,  $C_x X/N$ . This tax share can be transformed into  $s$ , the subjective marginal tax price per unit of  $q$  as

$$(5) \quad s = \frac{C_x X}{N} \cdot \frac{1}{q} = C_x N^{\alpha-1}$$

Again, for convenience, we posit the exact form of the median voter's demand schedule for  $q$  as log-linear in  $s$  and in  $y$ , the chooser's income.<sup>7</sup>

$$(6) \quad q = A s^\eta y^\delta$$

Using (3''), (4), and (5), this last equation can be rewritten in terms of per capita expenditure:<sup>8</sup>

$$(7) \quad e = A' w^{\beta(\eta+1)} y^\delta N^{(\alpha-1)(\eta+1)}$$

In logarithmic form this becomes

$$(7') \quad \ln e = \ln A' + (\eta + 1) \ln w^\beta + (\alpha - 1)(\eta + 1) \ln N + \delta \ln y$$

The above model has the two important properties that were earlier promised: it is consistent with the economic theory of majority rule and (with the addition of a random disturbance term) it permits estimation of the parameters that this theory suggests are most important. In particular, the model permits the estimation of the price elasticity of demand and the degree of publicness of the goods produced, parameters heretofore considered to be nonobservable.

## II. Estimation

In this section we shall estimate the parameters governing the demand for public services in the United States using the function developed in equation (7'). The empirical study is based on cross-section

<sup>7</sup> The reader may question the omission from (6) of the prices of other public services and other private goods. This omission was deliberate and based on two important considerations. First, had prices of other non-federal public services been included, the  $\alpha$ 's and  $\eta$ 's could have been estimated only by using a system of equations with non-linear constraints, a level of sophistication beyond the scope of our paper. Second, we omitted the price of all private goods because no suitable index now exists. Essentially this implies that the cross-effects of other public services are negligible and that the price of private goods is more-or-less the same across units.

<sup>8</sup> Equation (7) may be derived from the definition of per capita expenditure,  $e = sq$ .

tional data aggregated at the state level; the estimating year is 1962.<sup>9</sup> Demand functions were estimated for eight specific public services: local education, higher education, highways, health and hospitals, police, fire, sewers and sanitation, and parks and recreation.<sup>10</sup>

In order to estimate the demand functions we need a series of prices for the various services. Under our assumptions of elastic factor supply and *C-D* constant returns production, it was demonstrated that

$$(3'') \quad C_x = a'w^\beta$$

Estimation of  $\beta$  for each service category enables us to use wage data in deriving a series  $w^\beta$  which is proportional to marginal cost.<sup>11</sup>

The hypothesis of constant returns to scale is supported by evidence summarized by Hirsch (pp. 167-84). He cites empirical studies which indicate that average costs are constant over wide ranges of output for a large number of public services supplied by small, localized production units (e.g., police protection, local education, hospitals, sanitation, etc.) These findings together with cost-minimization and elas-

tic factor supplies within each unit imply a class of production functions exhibiting constant returns to scale.

The constancy of labor's share, a necessary condition for the *C-D* specification, also must be established. Random state-wise variation in labor's share (variation not related to output or wages) is expected and may imply either that some states are producing slightly different goods under the same classification (i.e., "highway services" in Montana may not be identical to highway services in New Jersey) or that some states use slightly different *C-D* functions than others to produce a homogeneous output. Neither of these two occurrences would necessarily violate the hypothesis that individual states use *C-D* techniques to supply public services. Only if labor's share varies systematically with either output or wage rates must we conclude that this *C-D* assumption is invalid.

The method used to test for the constancy of labor's share is equivalent to an analysis of variance procedure. For each public service classification, data were collected on total outlays and total payrolls of state and local governments for each of forty-four states<sup>12</sup> during the years 1960-64.<sup>13</sup> For each public service, a regression equation was fitted in which the dependent variable was the ratio of total payroll to total outlay (labor's share in each state in each year) and the independent

<sup>9</sup> The year 1962 was chosen for three reasons: 1) Census of Government data are available at five-year intervals (1957, 1962, 1967, . . . , etc); 2) Taxpayers appeared to be in a position of disequilibrium after 1963—a sort of "taxpayers revolution"—as evidenced by the high rates of bond issue failures. See Buchanan and Marilyn Flowers; 3) The labor quality index used was computed for the years 1954, 1958, and 1963 by Barzel (1967b).

<sup>10</sup> Other public services were not examined either because data were incomplete (as in the case of liquor control and utilities) or because the services were not final products (such as financial administration and general control).

<sup>11</sup> The assumptions about factor supplies were again dictated by convenience rather than strictly observable reality. Had data for actual and imputed rentals rates been available a more general supply function such as (3') or perhaps one employing a CES production function would have been used. Still, it seems intuitively plausible that rental rates do not differ over units, since capital markets are national in scope and transfer costs are likely to be negligible.

<sup>12</sup> Observations on these variables were gathered from various publications of the U.S. Census: Payrolls (1960-1964b) (for the month of October) and total expenditures (1960-64a). Data problems forced exclusion of six states and the District of Columbia; see fn. 17.

<sup>13</sup> Total cost data were not available and would differ from total outlay by an amount equal to the difference between purchases of new capital and implicit rental payments (opportunity costs) to owned capital. While total cost and total outlay may differ in any given year, cost minimization and competition guarantee that they will be equal over the long run. For this reason, we used data collected over a five-year interval in estimating labor's share.

variables were a constant, forty-three state dummy variables and four yearly dummies. Although the ranges of variation in labor's share were generally small, significant differences in state dummy coefficients were observed for some public service categories. When estimates of labor's share for the individual states (derived by adding state dummy coefficients to the constant term) were ordered by increasing magnitude of state outlay, no systematic relationship between labor's share and outlay was observed for any of the public service classifications. Similar attempts were made to relate variations in labor's share to wage rates with no further success. Thus, the variations appear to be random and do not contradict the assumption that *C-D* techniques are used to supply public services.

To ensure that prices and outputs (for each public service classification) are comparable among states, we decided to examine groups of states within which labor's share tended to be constant—i.e., states which appear to be using the same technology to produce a homogeneous output. Once these groups have been located we may estimate separate demand functions for different groups of states. For each public service category, the forty-three state dummy coefficients were plotted on the real line and clusters of coefficients which could be enclosed in an interval of two standard errors were noted. In four categories only one such group was noticed and hence only one demand function is estimated for each of these services; each of the other four services displayed two distinct groups. Within each group, the state with the median dummy coefficient was noted. The estimate of labor's share for this state (for 1962) was used as an estimate of  $\beta$ . Outlying states which did not fall within any group were deleted. Table 1 presents estimates of  $\beta$ , as well as ranges of variation of labor's share and the

TABLE 1—ESTIMATES OF  $\beta$  AND RANGES OF COEFFICIENT VARIATION WITHIN STATE GROUPS

	$\hat{\beta}^a$	Range of Variation within groups <sup>b</sup> $\pm$ standard error	Number of States in each group
Local Education;			
group 1	.7281	$\pm 0.8$ s.e.	19
group 2	.7899	$\pm 0.8$	14
Higher Education	.6449	$\pm 1.5$	31
Highways;			
group 1	.2104	$\pm 1.4$	24
group 2	.2647	$\pm 2.0$	19
Health-Hospitals	.5677	$\pm 2.0$	41
Police	.7947	$\pm 2.0$	43
Fire	.7487	$\pm 2.0$	43
Sewers-Sanitation;			
group 1	.2775	$\pm 1.0$	19
group 2	.4175	$\pm 1.5$	18
Parks-Recreation;			
group 1	.3489	$\pm 1.5$	23
group 2	.4485	$\pm 0.8$	17

Sources: See fn. 12.

<sup>a</sup> For each group,  $\hat{\beta}$  is the 1962 estimate of labor's share for the median state in that group.

<sup>b</sup> This is the maximum observed range of variation of state dummy coefficients within each group. This range is expressed in standard error units. Because there are an equal number of observations in each state category, the state dummy coefficients in any given equation all have a common standard error.

number of states within each group.

In addition to the data used in estimating  $\beta$ , observations were taken from each of the forty-four states (in 1962) on the following variables:<sup>14</sup> average wage rate of employees in each of the eight public service industries adjusted for quality (as explained below); average personal income of state residents;<sup>15</sup> state popula-

<sup>14</sup> All data were taken from Census publications: Average wage rates were derived from payrolls and full-time equivalent employment (1960-64b); expenditures (1960-64a); and average personal incomes, populations, land areas and degrees of urbanization (1961-1965).

<sup>15</sup> Federal per capita aid has been deliberately excluded from our model. If it were included as a separate variable, the income term would have to be reduced by the amount of federal taxes paid and increased by the value of the federal services received, i.e., an adjustment

tions; degree of urbanization (percentage of residents living in urban areas); and state land areas.

Since the marginal productivity of labor is unobservable, observed variation in wage rates may exist for one or both of two distinct reasons: quality differentials or mobility costs at the margins. The difference we wish to measure is the latter though observed wage differences could just as easily reflect quality differentials. For example, suppose we found that a teacher in state *A* received half the salary of a teacher in state *B*. Given the model this would suggest that the marginal cost of education in *B* was higher than in *A*. If, however, we were told that the quality of a teacher's service in *B* was twice that in *A*, the wages adjusted for quality would be equal. If an index of quality were available, we could adjust wages to a constant-quality labor unit and only mobility differentials would remain. Fortunately, Barzel has developed just such an index (for 1963)<sup>16</sup> and we have employed it in deflating our wage data.<sup>17</sup>

of income would have to be made for the federal "fiscal residuum." Since we do not know this residuum, it was felt unwise to include the aid. Thus, we arbitrarily assume that aid has no net wealth effect. There is also excellent reason to believe it has little price effect—even when tied to particular expenditures—since most grants are not open-ended and their effects are probably only inframarginal. This is discussed by David Bradford et al. We used average income as a surrogate for the income of our unobserved median preference holder. This procedure is justified under the assumption that the two quantities are highly correlated.

<sup>16</sup> Barzel (1967a, b) assumes that for any state, the quality of labor in the private services sector is directly proportional to the quality of labor in the manufacturing sector. Further, he argues that since the outputs, the capital and the firms in that latter sector are all very mobile, the observed differences in wages reflect quality differences alone. Due to a lack of data, the index was not computed for six states and the District of Columbia and these units were deleted from our study.

<sup>17</sup> Assume that  $L$ , labor in (1), is the product of two variables:  $l$  the observed unit of labor and  $h$ , a quality index for  $l$ . Since  $L = h \cdot l$ , we see from (2) that the true wage  $w = \beta \cdot C_x \cdot X / L = \beta \cdot C_x \cdot X / h \cdot l = \hat{w} / h$  where  $\hat{w}$  is the observed wage. Aside from the theoretical desirability of adjusting wages to reflect quality differentials, de-

The reader will note that we have introduced other variables not mentioned in the theoretical development of the model. Some explanation is in order. Although theory tells us that price, income, and population are important determinants of the demand schedule of the median voter, we suspect that this function may be systematically related to other measurable variables. For example, for a given price, income and population, one might expect per capita demand for highway services to be greater in states with large land areas than in smaller states.<sup>18</sup> Similarly, it seems plausible that per capita demand for fire protection may be greater in cities than in rural areas. While we have no formal theory to explain such differences in preferences, intuition suggests that the variables urbanization and state land areas may be relevant.

In light of these considerations, we decided to estimate four equations for each demand function:

$$(8a) \quad \ln e = \text{const.} + (\eta + 1) \ln w^\beta + \theta \ln N + \delta \ln y + \epsilon$$

$$(8b) \quad \ln e = \text{const.} + (\eta + 1) \ln w^\beta + \theta \ln N + \delta \ln y + \gamma_1 \ln \text{urbanization} + \epsilon'$$

$$(8c) \quad \ln e = \text{const.} + (\eta + 1) \ln w^\beta + \theta \ln N + \delta \ln y + \gamma_2 \ln \text{area} + \epsilon''$$

$$(8d) \quad \ln e = \text{const.} + (\eta + 1) \ln w^\beta + \theta \ln N + \delta \ln y + \gamma_1 \ln \text{urbanization} + \gamma_2 \ln \text{area} + \epsilon'''$$

flating by the labor quality index also diminished a rather strong correlation between income and unadjusted wage rates. Before adjustment, the simple correlations between income and wage rates ranged from .56 to .84; after adjustment, all were below .50.

<sup>18</sup> Because population is already in the demand equation, inclusion of state land areas allows per capita demand to be explained by population density.

where  $\theta = (\alpha - 1)(\eta + 1)$ .

Theory cannot indicate which of these four equations correctly specifies the demand for public services so two sets of results are reported. Table 2 displays results derived from estimating equation (8a). The results shown in Table 3 were obtained by including urbanization and/or area in the equation whenever their pres-

ence increased  $\bar{R}^2$  or produced parameter estimates which differed significantly (by one standard error or more) from the estimates of (8a).<sup>19</sup> Columns (3) and (4) in each table show estimated price and income elasticities, respectively (with stan-

<sup>19</sup> A justification of this procedure is given in Rao and Miller, pp. 35-38.

TABLE 2—REGRESSION RESULTS

	Parameter Estimates					$\bar{R}^2$ (6)
	const. (1)	$\hat{\theta}$ ( <i>ln pop</i> ) (2)	$\hat{\eta}$ price elasticity (3)	$\hat{\delta}$ income elasticity (4)	$\hat{\alpha}$ capturability parameter (5)	
Local Education						
group 1 ( $\beta = .7281$ )	-.8960 <i>1.6005</i>	-.0378 <i>.0347</i>	-1.1596 <i>.3398</i>	.8093 <i>.1701</i>	1.0349 <i>.0809</i>	.59
group 2 ( $\beta = .7899$ )	1.5441 <i>2.3154</i>	-.0133 <i>.0528</i>	-1.9018 <i>.7048</i>	.9522 <i>.2706</i>	1.0092* <i>.0366*</i>	.45
Higher Education ( $\beta = .6449$ )	-1.0922 <i>4.1134</i>	-.1818 <i>.0751</i>	-.1671 <i>.7801</i>	.2950 <i>.4307</i>	.8837* <i>.1183*</i>	.10
Highways						
group 1 ( $\beta = .2104$ )	.5624 <i>3.1874</i>	-.1459 <i>.0602</i>	2.3958 <i>2.3592</i>	-.0183 <i>.2698</i>	.8964* <i>.0844*</i>	.20
group 2 ( $\beta = .2647$ )	.2306 <i>1.7526</i>	-.2283 <i>.0404</i>	.4761 <i>.9154</i>	.3008 <i>.2322</i>	.6702* <i>.2124*</i>	.63
Health-Hospitals ( $\beta = .5677$ )	-.6312 <i>1.8537</i>	.0586 <i>.0431</i>	-1.1283 <i>.5713</i>	.4970 <i>.2245</i>	.9781 <i>.0989</i>	.11
Police ( $\beta = .7947$ )	-7.9530 <i>1.7354</i>	.1046 <i>.0310</i>	-.9001 <i>.3591</i>	1.2136 <i>.1370</i>	1.0752 <i>.0271</i>	.67
Fire ( $\beta = .7487$ )	-14.1946 <i>2.5240</i>	.0731 <i>.0469</i>	-.3203 <i>.5068</i>	1.6156 <i>.2040</i>	1.0691* <i>.0681*</i>	.61
Sewers-Sanitation						
group 1 ( $\beta = .2775$ )	-10.0609 <i>4.5832</i>	.1544 <i>.1081</i>	-1.4945 <i>2.5564</i>	.7326 <i>.8113</i>	.9887 <i>.0602</i>	.09
group 2 ( $\beta = .4175$ )	-12.3721 <i>2.6145</i>	.2607 <i>.0812</i>	-4.6576 <i>1.6668</i>	2.0059 <i>.4481</i>	.9402* <i>.0319*</i>	.56
Parks-Recreation						
group 1 ( $\beta = .3489$ )	-21.3659 <i>3.2597</i>	.2114 <i>.0919</i>	-.4958 <i>1.3683</i>	2.7359 <i>.4341</i>	1.0501 <i>.0312</i>	.68
group 2 ( $\beta = .4485$ )	-12.7187 <i>6.1638</i>	.0904 <i>.1104</i>	.4143 <i>1.9103</i>	1.2889 <i>.6042</i>	1.0723 <i>.1356</i>	.14

Sources: See fn. 14.

Note: Figures in italics are standard errors; Figures marked \* in column (5) were derived from estimates of  $\eta + 1$  where coefficient of variation was less than unity. These estimates are considered more reliable than the unstarred ones. See fn. 21.

TABLE 3—REGRESSION RESULTS

	Parameter Estimates							$\bar{R}^2$ (8)
	const. (1)	$\hat{\theta}$ ( $\ln$ pop) (2)	$\hat{\eta}$ price elasticity (3)	$\hat{\delta}$ income elasticity (4)	$\hat{\alpha}$ capturability parameter (5)	$\hat{\gamma}_1$ ( $\ln$ area) (6)	$\hat{\gamma}_2$ ( $\ln$ urban.) (7)	
Local Education	-1.9767	-.0382	-1.1276	.9385	1.0527	.0734		.68
group 1 ( $\beta = .7281$ )	<i>1.5154</i>	<i>.0310</i>	<i>.3047</i>	<i>.1633</i>	<i>.1220</i>	<i>.0337</i>		
group 2 ( $\beta = .7899$ )	-1.8690	-.1386	-1.2197	1.0422	1.0925	.1048		.74
	<i>1.8957</i>	<i>.0519</i>	<i>.5295</i>	<i>.1901</i>	<i>.2262</i>	<i>.0307</i>		
Higher Education	5.2378	-.2532	.0127	.6886	.8161*	.2310		.45
( $\beta = .6449$ )	<i>3.2893</i>	<i>.0607</i>	<i>.6078</i>	<i>.3479</i>	<i>.1191*</i>	<i>.0534</i>		
Highways	1.1061	-.2965	.5864	.1033	.9320	.1750	.3448	.42
group 1 ( $\beta = .2104$ )	<i>2.7803</i>	<i>.0800</i>	<i>2.0991</i>	<i>.2891</i>	<i>.0916</i>	<i>.0581</i>	<i>.2790</i>	
group 2 ( $\beta = .2647$ )	-.4871	-.2364	-.1752	.5354	.8685	.0718		.69
	<i>1.6383</i>	<i>.0371</i>	<i>.8960</i>	<i>.2420</i>	<i>.1440</i>	<i>.0358</i>		
Health-Hospitals	.3961	.0173	-1.1234	.1575	1.0065		.4166	.15
( $\beta = .5677$ )	<i>1.9164</i>	<i>.0491</i>	<i>.5581</i>	<i>.3010</i>	<i>.0352</i>		<i>.2528</i>	
Police	-6.4154	.0627	-.9691	.8154	1.0190		.4844	.75
( $\beta = .7947$ )	<i>1.5945</i>	<i>.0299</i>	<i>.3176</i>	<i>.1666</i>	<i>.1957</i>		<i>.1395</i>	
Fire	-11.7663	.0093	-.3543	.8799	1.0098*		.8735	.71
( $\beta = .7487$ )	<i>2.2636</i>	<i>.0436</i>	<i>.4364</i>	<i>.2605</i>	<i>.0474*</i>		<i>.2284</i>	
Sewers-Sanitation	-8.0656	.0778	-.8626	.0421	1.0017		.6254	.16
group 1 ( $\beta = .2775$ )	<i>4.6005</i>	<i>.1158</i>	<i>2.4923</i>	<i>.9055</i>	<i>.0310</i>		<i>.4172</i>	
group 2 ( $\beta = .4175$ )	-12.5297	.2759	-3.2450	1.5646	.9270*	-.1312		.59
	<i>2.5328</i>	<i>.0793</i>	<i>1.8554</i>	<i>.5376</i>	<i>.0915*</i>	<i>.0940</i>		
Parks-Recreation	-21.3659	.2114	-.4958	2.7359	1.0501			.68
group 1 ( $\beta = .3489$ )	<i>3.2597</i>	<i>.0919</i>	<i>1.3683</i>	<i>.4341</i>	<i>.0312</i>			
group 2 ( $\beta = .4485$ )	-9.1615	.0156	-.1765	.4902	1.0033		.8397	.18
	<i>6.6498</i>	<i>.1258</i>	<i>1.8771</i>	<i>.8646</i>	<i>.0269</i>		<i>.6637</i>	

Sources: See fn. 14.

Note: Figures in italics are standard errors; Figures marked \* in column (5) were derived from estimates of  $\eta+1$  whose coefficient of variation was less than unity. These estimates were considered more reliable than unstarred ones. See fn. 21.

dard errors in italics).<sup>20</sup> To aid the interpretation of results, we assume that the

<sup>20</sup> For purposes of estimation, per capita expenditures were deflated by the term  $w^\theta$  so that the dependent variable used was  $\ln(e/w^\theta)$ . This was done for convenience since it yields estimates of  $\eta$  (instead of  $\eta+1$ ) and  $t$ -ratios directly. It can be shown that estimates of the parameters and their standard deviations (from ordinary least squares) are unaffected by this transformation.

error terms follow a spherical normal distribution.

We estimate the capturability parameter  $\alpha$  using estimators of two other parameters,  $\eta$  and  $\theta$  (where  $\alpha = 1 + \theta/(\eta+1)$ ). Because the ratio of two unbiased estimators is not unbiased, an approximation (due to A. S. Merrill) is used. Approximations for the mean and variance of a ratio of two

variables  $(\theta/\eta+1)$  to the order  $T^{-1}$  are given by

$$(9) \quad \hat{\alpha} \cong 1 + \frac{\hat{\theta}/(\hat{\eta}+1)}{1 + \sigma_{\hat{\eta}}^2/(\eta+1)^2}$$

where

$$E(\hat{\alpha}) = 1 + \theta/(\eta+1) = \alpha$$

and

$$(10) \quad \text{Var}(\hat{\alpha}) \cong (\hat{\alpha}-1)^2 \{ \sigma_{\hat{\theta}}^2/\theta^2 + \sigma_{\hat{\eta}}^2/(\eta+1)^2 \}$$

where  $T$  is the sample size and we assume  $\hat{\theta}$  and  $\hat{\eta}$  are independent and

$$\hat{\theta} \sim N(\theta, \sigma_{\hat{\theta}}^2), \hat{\eta} \sim N(\eta, \sigma_{\hat{\eta}}^2)$$

Using the above formulae,<sup>21</sup> approximations of the capturability parameter and its standard deviation were computed. (Of course, we do not have population moments for the two variables  $\hat{\theta}$  and  $\hat{\eta}$ , and use unbiased estimates of these moments in all calculations.) Estimates of  $\alpha$  are presented in column (5) of Table 2 and Table 3 (with standard errors in parentheses).

### III. Conclusions

Comparison of the results of Tables 2 and 3 reveals that the inclusion of urbanization and area had little impact upon our parameter estimates. Although there are some significant differences in the two sets of estimates, the same general patterns are evident. Since our equations include population, the estimates of  $\gamma_1$  in Table 3,

column (6), show the partial effect of an increase in the dispersion of population upon per capita expenditure. This effect is positive and significant at the 5 percent level for local education, higher education, and highways—i.e., for these three services, per capita expenditure is inversely related to population density. This is, of course, the expected result for highways. Column (7) of Table 3 shows that urbanization is significant at the 5 percent level in only two equations—police and fire protection. In both cases, the degree of urbanization is positively related to per capita expenditure.

Income elasticities are not terribly different from those reported in other empirical studies, i.e., between +0.2 and +1.0. None of the estimates are significantly negative and fifteen are positive at a 5 percent level of significance.

Although the estimated price elasticities where statistically significant are all negative, estimates obtained for higher education and highways are frequently positive but uniformly insignificant, indicating that our measure of marginal cost is not a significant determinant of the demand for these services. Perhaps these findings reflect the fact that these two services are subsidized to a high degree by federal grants which bear no direct relationship to federal taxes paid by state residents. Hence, the connection between the tax price seen by voters and the true marginal cost of supplying these services may be rather tenuous.

Under the assumption that the true coefficient of variation of  $(\hat{\eta}+1)$  is less than .1, the distribution of  $\hat{\alpha}$  is approximately normal. Unfortunately, since sample coefficients of variation are in general greater than .1, we are not justified in making this assumption. Thus we cannot comment upon significance levels of individual estimates of  $\alpha$ . However, as noted in fn. 21 the starred estimates were derived from

<sup>21</sup> These formulations by Merrill are derived from a binomial expansion of the term  $\{1+\hat{\nu}/(\eta+1)\}^{-1}$  ( $\hat{\nu}$  is the deviation of  $\hat{\eta}$  from its mean). Because the binomial series converges only where  $|\hat{\nu}/(\eta+1)| < 1$ , these approximations are invalid when the coefficient of variation in the population of  $\hat{\eta}+1$  is large. To the extent that a high coefficient of variation in the population is reflected by a high estimate, we place relatively less reliance on approximations derived from high coefficient of variation estimates. The starred items in Tables 2 and 3 were derived from estimates of  $\eta+1$  whose coefficients of variation was less than unity and are considered more reliable than the unstarred items.

relatively precise estimates of  $(\eta+1)$  and are considered more reliable than the unstarred items.

In addition to these qualifications of our estimates, great care should be exercised in interpreting  $\alpha$  itself. Even in cases where  $\alpha$  is close to unity, there may exist substantial gains from collectivization of the expenditure decision. Thus, normative conclusions drawn from the finding that the goods appear better classified as private or quasi private rather than public are highly conjectural.

However, in spite of these caveats, the fact that most estimates of  $\alpha$  are greater than unity invites speculation. One explanation for this phenomenon lies in the net gains to be realized by a coalition formed around the median preference holder. If this group is able to secure legislation providing services which largely accrue to itself and/or differentially tax outsiders at higher rates, collectivization is efficient to this median group.<sup>22</sup> If this is the case, empirical knowledge of this discrimination would allow us to re-specify our model to allow for this purely political effect. Our omission of this possibility on the grounds of infeasibility only temporarily absolves us (and others in the field) from the task of accounting for this phenomenon.

#### REFERENCES

- K. J. Arrow, *Social Choice and Individual Values*, 2d ed., New York 1963.
- R. Barlow, "Efficiency Aspects of Local School Finance," *J. Polit. Econ.*, Sept./Oct. 1970, 78, 1028-39.
- J. L. Barr and O. A. Davis, "An Elementary Political and Economic Theory of Local Governments," *Southern Econ. J.*, Oct. 1966, 33, 149-65.
- Y. Barzel, (1967a) "The Price and Quantity of Services: The Use of Cross-Sectional Parameters in Time Series," unpublished 1967.
- , (1967b) "Wages in Manufacturing by State and Industry," unpublished 1967.
- W. C. Birdsall, "A Study of the Demand for Public Goods," in R. A. Musgrave, ed., *Essays in Fiscal Federalism*, Washington, 1965.
- D. Black, *The Theory of Committees and Elections*, Cambridge 1958.
- T. E. Borcharding, "The Growth of Non-Federal Public Employment in the United States, 1900 to 1963," unpublished doctoral dissertation, Duke Univ. 1966.
- H. R. Bowen, "The Interpretation of Voting in the Allocation of Economic Resources," *Quart. J. Econ.*, Nov. 1943, 58, 27-48.
- D. F. Bradford et al., Papers on the Economics of Political Decentralization in *Amer. Econ. Rev. Proc.*, May 1971, 61, 440-65.
- J. M. Buchanan and M. Flowers, "An Analytical Setting for a Taxpayer's Revolution," *Western Econ. J.*, Dec. 1969, 7, 349-59.
- J. M. Buchanan and G. Tullock, *The Calculus of Consent*, Ann Arbor 1962.
- R. H. Coase, "Discussion," of M. Abramovitz and V. Eliasberg, "The Trend of Public Employment in Great Britain and the United States," *Amer. Econ. Rev. Proc.*, May, 1953, 43, 234-36.
- O. A. Davis and G. H. Haines, Jr., "A Political Approach to a Theory of Public Expenditures: The Case of Municipalities," *Nat. Tax J.*, Sept. 1966, 19, 259-75.
- A. Downs, *An Economic Theory of Democracy*, New York 1957.
- W. L. Hansen and B. A. Weisbrod, *Benefits, Costs, and Finance of Public Higher Education*, Chicago 1969.
- W. Z. Hirsch, *The Economics of State and Local Government*, New York 1970.
- A. S. Merrill, "Frequency Distribution of an Index when Both the Components Follow

<sup>22</sup> George Stigler suggests that the median voting group (the economic middle class) has systematically shifted income towards itself in the political process. W. Lee Hansen and Burton A. Weisbrod indicate that for higher education this is true. David G. Tuerck, concludes that constitutional constraints severely impeded arbitrary treatment of taxpayers, but no such limitation is imposed upon the expenditures side.

- the Normal Law," *Biometrika*, July 1928, 20A, 53-63.
- P. Rao and R. L. Miller, *Applied Econometrics*, Belmont 1971.
- G. J. Stigler, "Director's Law of Public Income Distribution," *J. Law Econ.*, Apr. 1970, 13, 1-10.
- D. G. Tuerck, "Constitutional Asymmetry," *Publ. Choice*, 2, 1967, 27-44.
- G. Tullock, *Towards a Mathematics of Politics*, Ann Arbor 1967.
- U.S. Bureau of Census, *Census of Government, 1962: Compendium of Government Finances*, Washington 1963.
- , (1960-64a) *Governmental Finances*, Washington 1960-64.
- , (1960-64b) *State Distribution of Public Employment*, Washington 1960-64.
- , *Statistical Abstract of the United States*, Washington 1961-65.

# The Prediction Performance of the FRB-MIT-PENN Model of the U.S. Economy

By CHARLES R. NELSON\*

This paper presents an evaluation of the prediction performance of the FRB-MIT-PENN (*FMP*) econometric model of the U.S. economy using predictions provided by simple time-series models to establish standards of accuracy. The motivation for such an evaluation is two-fold. First, the quality of predictions provides a measure of the success of the model in simulating the behavior of the system under study. Second, we may be interested in the utility of the model for operational forecasting and policy design. The continuing development of the *FMP* model has resulted in a succession of revisions, the subject of this study being Version 4.1 which was released during 1969.<sup>1</sup> It should be emphasized that this evaluation is not intended as an audit of current developmental efforts, rather the objective is to consider a version of the model which has been thoroughly checked for stability and computational anomalies and which has been the subject of considerable interest and research in academic, corporate, and governmental policy contexts.

The study focuses on the one-quarter-

ahead predictions of fourteen endogenous variables of general interest; namely nominal *GNP*, its endogenous components, the unemployment rate, two price indices, and three interest rates. Predictions are obtained from the model by simultaneous solution of the equation system, requiring as inputs the historical values of endogenous and exogenous variables and projected *future* values of exogenous variables. In order to avoid ambiguity with regard to a method for projecting exogenous variables, these were set at their *actual* future values. These actual values provide *ex post* information to the model which is exploited by the behavioral relationships and provides some of the accounting components of *GNP*. Such predictions may be thought of as forecasts which would have been made by a user of the model who was endowed with perfect foresight with regard to future values of exogenous variables.

The computation of predictions from the *FMP* model amounts to evaluation of the conditional expectations of future endogenous variables implied by the equation system, since future values of the stochastic disturbances in the system are set at their expected values of zero.<sup>2</sup> If the

\* Assistant professor, Graduate School of Business, University of Chicago. I am grateful to J. Phillip Cooper, Eugene Fama, John P. Gould, P. Tinsley, Victor Zarnowitz, and particularly to Harry Roberts and Arnold Zellner for helpful comments, but, of course, retain responsibility for errors. This study was supported in part by a grant from the First National Bank of Chicago.

<sup>1</sup> Version 4.1 has been described by Arnold Zellner (1969). The *FMP* model has also been described by Franco Modigliani, Robert Rasche, and J. Phillip Cooper; Frank DeLeeuw and Edward Gramlich; and Rasche and Harold Shapiro.

<sup>2</sup> Solution values so obtained may differ from conditional expectations somewhat due to the nonlinearity of the system. However, explicit reduced form solution of the system is infeasible and we must assume that departures due to disturbances entering into non-linear relationships are relatively minor. For evidence of the approximate linearity of the model, see Zellner and Stephen Peck.

observed data being predicted were generated by the *FMP* system, then the conditional expectation predictions being computed would constitute minimum mean square error predictions amongst those conditioned on the same set of information.<sup>3</sup> In this case, that information set includes historical data on all of the endogenous and exogenous variables of the system, as well as future values of exogenous variables.

The time-series models used in this study to establish standards of accuracy for the *FMP* model are empirical representations of individual endogenous variables as stochastic processes of integrated autoregressive moving average (*ARIMA*) form, following the methodology developed by George E. P. Box and Gwilym M. Jenkins. Given a model for a particular series, predictions are obtained by computing the expected values of future observations which are implied by the model conditional only on the past history of the series. The fact that the information set utilized by the *FMP* model (that is the histories of *all* variables in the system as well as actual future values of exogenous variables) subsumes the set available to the *ARIMA* models (just the past history of the variable being forecast) motivates the choice of the time-series models as standards of accuracy. To the extent that the economy behaves "as if" it were being generated by the *FMP* model, then the larger information set used by the model should be useful in reducing mean square prediction error relative to the *ARIMA*

models. Further, if *FMP* and *ARIMA* predictions are combined in a composite prediction to minimize mean square error, we would expect little contribution from the *ARIMA* predictions. On the other hand, if *FMP* predictions prove to be relatively inaccurate and *ARIMA* predictions contribute substantially to a composite prediction, then we would be led to conclude that the *FMP* model had underutilized the information available to it, presumably because of statistical and economic errors of specification and sampling errors in parameter estimates.

Previous authors in the area of prediction evaluation have frequently used "naive" prediction models to obtain standards of accuracy (for example, see Geoffrey Moore, H. O. Steckler, Victor Zarnowitz). The *ARIMA* models used in this study may be viewed as "not-quite-so-naive" models. While they indeed are economically naive (for example, predictions are not constrained to satisfy accounting identities), they are based on statistically sophisticated analysis. The reason for preferring the *ARIMA* models over more naive schemes is that their implementation is founded on statistical theory and places particular emphasis on selection of models appropriate to the stochastic structure of individual time-series. If the fitted models are appropriate representations of the stochastic structures of the variables being predicted, then the implied conditional expectations of future observations will in general provide more accurate predictions than will naive models of arbitrary form.

The results of comparison of *FMP* and *ARIMA* model prediction accuracy reported in this study indicate that the former were more accurate for most of the variables during the sample period over which both models were fitted. When the two sets of predictions are combined into linear composites, most of the *ARIMA*

<sup>3</sup> To see this, denote by  $F$  a prediction of actual value  $A$  which is a function of information set  $\Phi$  and differs from the conditional expectation  $E(A|\Phi)$  by amount  $d$ . The expectation of the squared error  $(A-F)$  is then

$$\begin{aligned} E[(A-F)^2|\Phi] &= E\{[A-E(A|\Phi)]^2|\Phi\} \\ &\quad + 2dE\{[A-E(A|\Phi)]|\Phi\} + d^2 \\ &= E\{[A-E(A|\Phi)]^2|\Phi\} + d^2 \end{aligned}$$

so that mean square error is at a minimum when  $d=0$ ; that is, when the prediction  $F$  is set at  $E(A|\Phi)$ .

predictions make a statistically significant contribution to accuracy over the sample period. Further, when composites are constructed to minimize joint loss across variables, the *ARIMA* models make significant contributions for almost all variables. Examination of post-sample errors suggests, however, that the *ARIMA* models are relatively more robust with respect to prediction outside the sample period. Among the alternatives of *FMP*, *ARIMA*, and composite predictions, the *ARIMA* predictions achieve the smallest mean square error for seven of fourteen variables, the composites for five, and the *FMP* predictions for only two.

### I. *ARIMA* Models for Endogenous Variables of the *FMP* Model

The class of time-series models under consideration for comparison with the *FMP* model is that for which some difference of the observed series may be represented as a stationary stochastic process of autoregressive moving average forms. Letting  $z_t$  denote the observed value of series  $z$  at time  $t$  and  $B$  denote the backshift operator (i.e.,  $B^k z_t = z_{t-k}$ ) then the sequence  $\{z_t\}$  is said to have a representation as an *ARIMA* process if  $z_t$  may be expressed as

$$(1) \quad \phi_p(B)(1 - B)^d z_t = \theta_0 + \theta_q(B)a_t$$

where  $\phi(B)$  and  $\theta(B)$  are polynomials in the backshift operator of degrees  $p$  and  $q$ , respectively, having zeros outside the unit circle,  $\theta_0$  is a constant, and  $\{a_t\}$  is a sequence of random disturbances. Thus, the  $d$ th difference of the observed series is a  $p$ th order autoregression with a  $q$ th order moving average disturbance and is both stationary and invertible. The appeal of this class of processes as empirical models derives from their compatibility with a very wide range of autocorrelation structures and hence a very wide range of stationary and nonstationary behavior.

Empirical application begins with what Box and Jenkins refer to as "identification," that is, specification of dimensions  $p$ ,  $d$ , and  $q$  on the basis of sample autocorrelations and partial autocorrelations of successive differences of the raw data. Parameters of an identified model are fitted by iterative minimization of the sum of squared residuals,  $\sum \hat{a}_t^2$ , which provides maximum likelihood estimates under the assumption that disturbances are normal. Various diagnostic procedures, particularly residual analysis, are applied to check the adequacy of the model. Given the fitted model, predictions of any desired horizon are computed by direct evaluation of the expected values of the future observations being predicted conditional only on the past history of the individual series.<sup>4</sup>

As in linear regression, the sum of squared residuals and therefore of one-step-ahead prediction errors may be reduced over the period of fit simply by addition of more "independent variables," that is, more autoregressive or moving average terms. The criterion of model selection followed in this study has been the representation of each series in the most parsimonious form which is consistent with its stochastic structure. The parameters included in the models are those for which estimates are significant or which are required to eliminate serial correlation in residuals. Thus the procedure has not been to minimize the variance of prediction errors over the general class of *ARIMA* models, rather it has been to obtain the simplest adequate representations. It has been my objective to apply the methodology in the most straightforward fashion so that these models would presumably be duplicated, except

<sup>4</sup> For a summary of the Box and Jenkins methodology including an illustrative application, the reader is referred to chs. 2 and 5 of my study, *The Term Structure of Interest Rates*.

for minor differences, by another investigator. This objective required that prior information of various sorts be disregarded. For example, *logs* rather than levels of output variables presumably exhibit spatial stochastic homogeneity. *Logs* were not used, however, since the postwar data to which the study was confined do not by themselves provide strong evidence against homogeneity in the raw levels. Also, inclusion of a constant term in models for interest rates would clearly have improved both sample and post-sample period performance. These terms were omitted, however, because they were not, interestingly enough, statistically significant.

Most equations of Version 4.1 were estimated through 1966-04. In order to maintain comparability with respect to data base, the *ARIMA* models were also estimated through 1966-04. Models for the fourteen endogenous variables included in the study are displayed in the Appendix. It suffices to note that the models generally involve rather few parameters and few lagged values. The range of models represented is quite broad, including both pure autoregressive and pure moving average models as well as mixed models. An interesting by-product of fitting the models was evidence from the autocorrelations of residuals of a *negatively* seasonal component in some of the standard seasonally adjusted series. In the cases of consumer expenditures on non-durable goods, housing expenditures, and the *GNP* deflator and to a lesser extent *GNP* itself, the residual  $\hat{a}_t$  for a given quarter tended to be negatively related to the residual appearing four quarters later. The implication of this finding is that the seasonal adjustment procedures in general use may "overadjust" series with particular stochastic structures and thereby introduce a negative seasonal relationship. This would tend to reinforce the idea

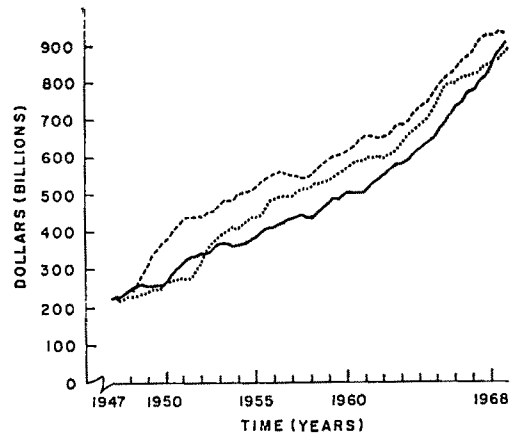


FIGURE 1. NOMINAL GROSS NATIONAL PRODUCT 1947-01 THROUGH 1969-01: HISTORICAL VALUES (SOLID LINE) AND TWO SIMULATIONS OF *ARIMA* MODEL

that unadjusted data should be utilized wherever possible and that seasonality should be accommodated by the econometric or statistical model under construction.<sup>5</sup>

The *ARIMA* model for nominal *GNP* is probably of special interest and is of remarkably simple form, namely

$$(2) \quad GNP_t - GNP_{t-1} = .615(GNP_{t-1} - GNP_{t-2}) + 2.76 + a_t$$

$$\hat{\sigma}_a^2 = 22.9$$

Thus, the change in the current quarter is simply related to the change in the previous quarter. Artificial realizations of postwar *GNP* were generated with the model by drawing random  $a_t$ s having the same variance as that of the residuals from estimation and using the initial quarters 1947-01 and 1947-02 as starting values. These simulations appear in Figure 1 along with the historical record through 1969-01. The familiar features of eco-

<sup>5</sup> The problems of seasonal adjustment and interpretation of adjustment procedures have been greatly illuminated by the recent work of David M. Grether and Marc Nerlove.

nomie history, recessions and booms, are easily recognized in the simulations and indistinguishable in character from actual episodes.<sup>6</sup>

## II. Analysis of Sample Period Prediction Errors

The system of equations of the *FMP* model requires a substantial number of lagged observations for solution, so that 1956-01 is the first quarter for which one-quarter-ahead predictions may be computed. Housing expenditures are exogenous through 1956 and only become endogenous in 1957 providing a more restricted prediction sample that begins in 1957-01. Thus, while we shall refer to 1956-01 through 1966-04 as the "sample period," both the *FMP* equations and the *ARIMA* models were generally fitted over periods that began before 1956-01. The analysis is confined to one-quarter-ahead predictions although predictions of longer horizon are available from both the *FMP* and *ARIMA* models and are of considerable practical interest. The intention is to concentrate on a more thorough analysis of the one-step-ahead case than would be tractable if a wider range of horizons were included. Hopefully, our qualitative conclusions remain valid for multiperiod prediction.

The mean squares, means, and standard deviations of sample period prediction errors appear in Table 1 and indicate that the *FMP* model provided generally more accurate predictions during that 44-quarter interval, although the differences are surprisingly small. In the cases of State and Local Government Expenditure and the Unemployment Rate, the *ARIMA* model predictions had smaller mean square errors. Mean errors were all small,

<sup>6</sup> For an early demonstration that random disturbances can account for cyclical phenomena, see Eugen Slutsky. Also see Ragnar Frisch. Stochastic simulations of the *FMP* model have been described by Cooper and Stanley Fischer.

suggesting that the prediction bias may be characterized as being minor. To investigate further the question of prediction bias, the actual values of endogenous variables were regressed on predictions. The predictions are properly regarded as the independent variables in these regressions since if they were correlated with their respective error terms then their prediction accuracy could be improved merely by exploitation of that correlation.<sup>7</sup> The constant term was significant (at the 5 percent level) only in the case of *FMP* predictions of State and Local Government Expenditures for which it was \$1.54 billion. Estimated slopes were significantly different from unity only for *FMP* predictions of Expenditures on Producers' Structures (1.04) and State and Local Government Expenditures (.97) and for *ARIMA* predictions of Expenditures on Producers' Durables (1.05). While these deviations from theoretical values are statistically significant, they are of rather small magnitude, thus reinforcing our previous conclusion that prediction bias is fairly small.

The correlation between *FMP* and *ARIMA* errors provides a measure of similarity between the two sets and is substantial for many of the variables including *GNP*. The highest error correlations are for Consumers' Expenditures on Non-durable Goods, Nonfarm Inventory Investment, and Expenditure on Producers' Structures. The lowest correlations are for State and Local Government Expenditures, where the *FMP* predictions did poorly, and for Yields on U.S. Treasury Bills.

Among desirable properties for predictions are that their errors be successively uncorrelated and that the predictions themselves be uncorrelated with future

<sup>7</sup> As John Muth pointed out this must be a property of rational expectations. Further, this must be a property of all conditional expectation predictions.

errors. Failure to meet either of these conditions implies underutilization of information, that is, predictions may be adjusted to reduce mean square error. If one-step-ahead errors are serially correlated, then predictions could have been improved upon by simply taking into account the relation between past and future errors. If predictions are correlated with corresponding future errors, then this relationship may be used to adjust predictions accordingly. Sample correlations between predictions and errors reported in Table 1 are small except for *FRB* predictions of Expenditures on Producers' Durables and State and Local Government Expenditures for which the mean square error was high. The sample correlation is also substantial for *ARIMA* predictions of Expenditures on Producers' Durables. Sample autocorrelations of *FMP* errors are large at lag one quarter (relative to a standard error of .16 for sample autocorrelations of uncorrelated noise) for variables, Expenditures on Producers' Durables, Expenditures on Producers' Structures, State and Local Government Expenditures, Housing Expenditures, and the *GNP* Deflator. Results for the Consumer Goods Price Index shows substantial correlation of *FMP* errors at longer lags (3 and 4 quarters). *ARIMA* model errors are relatively less autocorrelated but display strong correlation at lags 2 and 4 for the *GNP* Deflator. By way of summarizing these results for the two sets of prediction errors, we note that seven of the fifty-six autocorrelations for *FMP* errors lie outside the bounds  $\pm .32$  compared to one value of .32 for *ARIMA* model errors.

From the viewpoint of the operational forecaster, relationships between errors for different variables are important in prediction evaluation since his loss function will depend in general on such relationships. For example, in the case that

his loss function is a quadratic form in the prediction errors, as we assume in Section IV, then expected loss is a weighted sum of covariances between errors as well as individual error variances. Correlations across variables for both *FMP* and *ARIMA* errors appear in Table 2. A large correlation between *ARIMA* errors for a pair of variables would suggest that factors accounting for their respective contemporaneous disturbances are common to both. Thus, it is not surprising to find substantial positive correlation between *GNP* errors (disturbances) and those for Consumers' Expenditures on Durable Goods, Nonfarm Inventory Investment, and Expenditures on Producers' Durables, and substantial negative correlation with those for the Unemployment Rate. Errors for Expenditures in Producers' Durables are strongly correlated with those for Consumers' Expenditures on Durable Goods and Nonfarm Inventory Investment. Perhaps surprisingly, these three investment categories show quite strong positive correlation with the disturbances in the three interest rate series.

Correlations between contemporary errors of *FMP* predictions are generally indicative of the structure of the model, and, of course, of accounting relationships. Consequently, in Table 2 errors for *GNP* are positively related to those for its components, and negatively to the Unemployment Rate errors. Errors for Housing Expenditures and Expenditures on Producers' Structures are positively related. Where relationships are not so obvious on prior grounds, the correlations provide indications of structural interaction within the model, and may help to locate problem areas. For example, it is surprising that *GNP* Deflator and Consumer Price Index errors are negatively related to those for *GNP* and Expenditures on Producers' Durables. It is also interesting that errors for the three interest rates are positively

TABLE 1—SUMMARY STATISTICS FOR *FMP* MODEL AND *ARIMA* MODEL SAMPLE PERIOD PREDICTION ERRORS

Endogenous Variable	<i>FMP</i> Model Errors <sup>a</sup>			<i>ARIMA</i> Model Errors <sup>a</sup>			Correlation Between Model Errors and <i>ARIMA</i> Model Errors
	<i>MSE</i>	Mean	Standard Deviation	<i>MSE</i>	Mean	Standard Deviation	
1. <i>Gross National Product</i>	11.565	.695	3.344	25.330	.520	5.006	.468
2. Consumers' Expenditures on Nondurable Goods	1.926	.008	1.388	2.671	.230	1.618	.673
3. Consumers' Expenditures on Durable Goods	1.278	-.034	1.130	3.258	.057	1.804	.581
4. Nonfarm Inventory Investment	6.710	.892	2.432	10.992	.376	3.294	.640
5. Expenditures on Producers' Durables	.582	-.224	.795	1.050	.116	1.018	.238
6. Expenditures on Producers' Structures	.249	.064	.495	.310	.013	.557	.627
7. State and Local Government Expenditures	.570	.018	.755	.294	.136	.525	.121
8. Housing Expenditures <sup>b</sup>	.206	.092	.445	.486	.006	.697	.297
9. Unemployment Rate	.134	-.087	.356	.089	.040	.296	.432
10. <i>GNP</i> Deflator-Price Index	.338	.022	.194	.053	.004	.230	.402
11. Consumer Goods Price Index	.352	.018	.227	.062	.028	.247	.321
12. Yields on U.S. Treasury Bills	.361	.012	.247	.124	.015	.352	.173
13. Yields on Commercial Paper	.353	.044	.227	.101	.048	.314	.324
14. Yield on Corporate Bonds	.310	.020	.097	.016	.034	.120	.568

Endogenous Variable	Correlation Between Predictions and Errors		Serial Correlation Coefficients of Prediction Errors <sup>c</sup>							
			<i>FMP</i>				<i>ARIMA</i>			
	<i>FMP</i>	<i>ARIMA</i>	<i>r</i> <sub>1</sub>	<i>r</i> <sub>2</sub>	<i>r</i> <sub>3</sub>	<i>r</i> <sub>4</sub>	<i>r</i> <sub>1</sub>	<i>r</i> <sub>2</sub>	<i>r</i> <sub>3</sub>	<i>r</i> <sub>4</sub>
1. <i>Gross National Product</i>	.200	.263	.09	-.04	.05	.12	-.06	-.03	.05	-.14
2. Consumers' Expenditure on Nondurable Goods	.055	.174	-.19	.04	-.07	-.07	-.13	.00	.13	-.09
3. Consumers' Expenditure on Durable Goods	.005	.161	-.21	.17	-.08	.00	-.09	.12	-.07	.06
4. Nonfarm Inventory Investment	.191	-.096	.11	.01	-.08	-.12	.08	-.21	-.01	.04
5. Expenditure on Producers' Durables	-.152	.406	.38	.15	-.01	.14	.18	.25	-.07	.08
6. Expenditure on Producers' Structures	.303	.032	.21	.26	.28	-.15	-.13	.02	.08	.05
7. State and Local Government Expenditures	-.555	.224	.44	.43	.44	.35	-.09	-.10	.05	.03
8. Housing Expenditures <sup>b</sup>	.216	-.224	.34	.19	.24	.19	.09	-.01	.06	-.04
9. Unemployment Rate	.120	-.039	.01	.12	-.08	-.14	-.03	-.03	-.07	-.16
10. <i>GNP</i> Deflator-Price Index	-.160	-.108	.31	.16	.24	.31	-.09	.28	-.05	.32
11. Consumer Goods Price Index	-.199	-.132	.10	.11	.25	.33	.00	.24	.09	.23
12. Yields on U.S. Treasury Bills	-.058	-.179	-.23	-.20	.05	.09	-.01	-.11	-.17	.23
13. Yields on Commercial Paper	.080	-.165	.07	-.18	.01	.15	.04	-.21	-.09	.27
14. Yield on Corporate Bonds	-.067	-.204	.15	-.08	.10	.04	-.01	-.17	.01	.17

<sup>a</sup> Errors are in billions of current dollars for variables 1-8, and percentage points for the remaining variables.

<sup>b</sup> Sample period for Housing Expenditures is 1957-01 through 1966-04.

<sup>c</sup> The estimated standard error of *r*<sub>1</sub> under the hypothesis that the errors are uncorrelated is .16.

related to errors for Consumers' Expenditures on Durable Goods, Nonfarm Inventory Investment, and Expenditures on Producers' Durables. These correlations may be indicative of the origin of shocks in the real sector and their impact on the financial sector. They may also, of course, be indicative of problems in the structure of the model. The chains of interaction in a model of this size are enormously com-

plex and examination of error correlations may facilitate otherwise unwieldy diagnostic analysis.

The reader may have noted by this point the absence of analysis of "turning point" errors, a topic which has practically become standard in analysis of prediction accuracy (see previous references as well as Henri Theil). The argument for the importance of turning point errors, that

TABLE 2—CORRELATION OF SAMPLE PERIOD PREDICTION ERRORS ACROSS VARIABLES:  
FMP ERRORS ABOVE THE DIAGONAL, ARIMA BELOW THE DIAGONAL

Endogenous Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. <i>Gross National Product</i>	1.00	.20	.38	.74	.55	.28	.14	.34	-.35	-.16	-.27	.23	.33	.51
2. Consumers' Expenditures on Nondurable Goods	.13	1.00	-.16	-.19	-.06	.08	-.08	.14	.03	.27	.28	-.09	-.07	.04
3. Consumers' Expenditures on Durable Goods	.48	.09	1.00	.04	.22	.23	.03	.22	-.30	-.20	-.32	.22	.11	.30
4. Nonfarm Inventory Investment	.66	-.03	.18	1.00	.26	.12	-.14	.28	-.32	-.05	-.16	.22	.32	.38
5. Expenditures on Producers Durables	.50	.18	.63	.35	1.00	-.06	.37	.11	.09	-.41	-.37	.02	.20	.26
6. Expenditures on Producers' Structures	.20	.04	.25	.19	.12	1.00	-.19	.35	-.22	-.00	-.02	-.05	-.00	-.02
7. State and Local Government	.06	.11	.13	.14	-.02	.12	1.00	.16	-.01	-.37	-.28	-.08	-.13	-.02
8. Housing Expenditures <sup>a</sup>	.33	-.12	.30	.12	.03	.24	-.18	1.00	-.42	-.10	-.15	.03	.10	.05
9. Unemployment Rate	-.62	-.17	-.32	-.60	-.37	-.41	-.12	-.18	1.00	-.14	-.07	-.31	-.20	-.18
10. <i>GNP</i> Deflator-Price Index	-.05	-.06	-.03	.08	.06	.13	.21	-.13	-.23	1.00	.95	.16	.06	-.07
11. Consumer Goods Price Index	-.33	-.20	-.29	-.00	-.08	.05	.16	-.42	.05	.65	1.00	-.00	-.09	-.29
12. Yield on U.S. Treasury Bills	.34	.29	.30	.26	.40	.11	-.22	.19	-.30	-.22	-.31	1.00	.86	.53
13. Yield on Commercial Paper	.34	.20	.24	.44	.45	.12	-.17	.12	-.39	-.06	-.10	.89	1.00	.65
14. Yield on Corporate Bonds	.40	.29	.16	.33	.40	.11	-.21	.13	-.40	-.05	-.10	.77	.74	1.00

<sup>a</sup> Sample period for Housing Expenditures entries is 1957-01 through 1966-04.

is, errors in predicting the direction of change, has been well stated by Zarnowitz, p. 51, and, briefly, goes as follows. Economic time-series show strong systematic movements—trends and cycles. It should, then, be relatively easy to predict the continuation of a rise or fall. Consequently, to predict the end of the current movement and the beginning of the next appears to be a more crucial goal.

The crucial element in the argument for the importance of turning points is the view that cycles and trends in economic time-series are *systematic*. However, as we have seen from simulations of the *ARIMA* representations of *GNP*, "cycles" are not necessarily systematic in nature, but rather may be merely artifacts of random shocks working their way through the economy as Slutsky and Frisch suggested some time ago. Thus, it appears *ex post* that if turning points had been foreseen, prediction errors for subsequent observations could have been reduced. Turning points are usually associated with the occurrence of unusually large shocks to the system, and presumably success in an-

tipicating *any* large disturbance would contribute to the accuracy of predictions of subsequent observations. If that is the case, then we should not restrict our attention only to the large disturbances which produce turning points, but rather should be interested in anticipation of all large disturbances. In other words, to say that turning points are important because they are difficult to predict is only to say that large disturbances are associated with large prediction errors. Statistical decision theory offers further clarification on this point. Namely, once the loss associated with errors have been specified, then conditions for optimal predictions may be stated, as for example in minimization of mean square error. Thereafter, turning point errors are of no special interest in and of themselves.

### III. Composite Predictions of Endogenous Variables

Predictions computed from the *FMP* model are essentially the conditional expectations of future realizations implied by the structure of the model and the in-

$$(5) \quad \hat{\beta} = \frac{\sum [(FMP)_t - (ARIMA)_t][A_t - (ARIMA)_t]}{\sum [(FMP)_t - (ARIMA)_t]^2}$$

formation set available to it. If the *FMP* model make efficient use of that information, that is, if it does in fact provide conditional expectation predictions, then the *ARIMA* models which draw on only a subset of the same information should not be able to contribute to the accuracy of composite predictions which combine both.

A linear composite prediction is of the form

$$(3) \quad A_t = \beta_1(FMP)_t + \beta_2(ARIMA)_t + \epsilon_t$$

where  $A_t$  denotes the actual value for period  $t$ ,  $\beta_1$  and  $\beta_2$  are fixed coefficients, and  $\epsilon_t$  is the composite prediction error. Least squares fitting of (3) requires minimization of  $\sum \epsilon_t^2$  over values of  $\beta_1$  and  $\beta_2$  and therefore provides the minimum mean square error linear composite prediction for the sample period. In the case that both *FMP* and *ARIMA* predictions are individually unbiased, then (3) may be rewritten simply as

$$(4) \quad A_t = \beta(FMP)_t + (1-\beta)(ARIMA)_t + \epsilon_t$$

The least squares estimate of  $\beta$  in (4) is then given by equation (5) which is seen to be the coefficient of the regression of *ARIMA* prediction errors,  $[A_t - (ARIMA)_t]$ , on the difference between the two predictions. As would seem quite reasonable, the greater the ability of the difference between the two predictions to account for errors committed by  $(ARIMA)_t$ , the larger will be the weight given to  $(FMP)_t$ .

Consider now the hypothetical case that the *FMP* predictions subsume the *ARIMA* predictions and contain additional information  $(FMP')_t$  so that

$$(6) \quad (FMP)_t = (ARIMA)_t + (FMP')_t$$

Then

$$(7) \quad \hat{\beta} = \frac{\sum (FMP')_t [A_t - (FMP)_t + (FMP')_t]}{\sum (FMP')_t^2}$$

and it is readily seen that

$$(8) \quad \text{Plim } \hat{\beta} = 1$$

since *FMP* predictions are presumably uncorrelated with their associated errors. Thus, if *FMP* predictions do incorporate all of the information provided by *ARIMA* predictions, then estimates  $\hat{\beta}_1$  and  $\hat{\beta}_2$  in (3) should be approximately unity and zero, respectively.

Further insight into the structure of composite predictions is provided by an analogy to asset portfolios, namely composite predictions may be viewed as "portfolios" of predictions. If we denote individual *FMP* and *ARIMA* errors by  $u_1$  and  $u_2$ , respectively, then from (4) the composite prediction error is seen to be

$$(9) \quad \epsilon_t = \beta(u_{1t}) + (1-\beta)(u_{2t})$$

Thus, just as the return on a portfolio of assets is the weighted average of individual asset returns, the composite error is the weighted average of individual errors. In both cases, the objective is to minimize the variance of the weighted average given its expected value.<sup>8</sup> Construction of efficient asset portfolios requires selection of weights that minimize variance for various values of expected return, while in the case of prediction portfolios the

<sup>8</sup> In the case of two assets this is trivial since specification of a given rate of return determines both weights (except when the expected returns on both assets are identical) and leaves no additional degrees of freedom. The conceptual analogy, however, generalizes to many assets and many predictions.

weighted average always has expectation zero if individual predictions are unbiased or may be given expectation zero by addition of an appropriate constant.

Minimizing composite error variance over a finite sample of observations leads to the estimate of  $\beta$  given by

$$(10) \quad \hat{\beta} = \frac{\sum u_{2t}^2 - \sum u_{1t}u_{2t}}{\sum u_{1t}^2 + \sum u_{2t}^2 - 2\sum u_{1t} \cdot u_{2t}}$$

For large samples, or in the case that the variances  $V(u_{1t})$  and  $V(u_{2t})$  and the covariance  $C(u_{1t}, u_{2t})$  are known, we have

$$(11) \quad \beta = \frac{V(u_{2t}) - C(u_{1t}, u_{2t})}{V(u_{1t}) + V(u_{2t}) - 2C(u_{1t}, u_{2t})}$$

Thus, the minimum variance weight  $\beta$  is seen to depend on the covariance between individual errors as well as on their respective variances, just as the analogous weight for a minimum variance two-asset portfolio depends on the covariance of returns as well as on return variances. Holding covariance constant, the larger the variance of the *ARIMA* error relative to that of the *FMP* error, the larger the weight given to the *FMP* prediction. However, with the exception of the special case  $C(u_{1t}, u_{2t}) = V(u_{1t})$ ,  $\beta$  will always differ from unity and therefore the weight given to the *ARIMA* prediction differ from zero no matter how extreme might be the ratio of their error variances. *In short, relative accuracy is not an appropriate basis for choosing one prediction to the exclusion of the other; rather, even a very inaccurate prediction would generally be included in a minimum variance composite.*

Considering again the limiting case where the *FMP* prediction subsumes all the information in the *ARIMA* prediction, we see from expression (6) that the *ARIMA* error  $u_{2t}$  would be given by

$$(12) \quad u_{2t} = u_{1t} + (FMP')_t$$

Since errors are presumably uncorrelated with corresponding predictions, we have

$$(13) \quad V(u_{2t}) = V(u_{1t}) + V(FMP')_t$$

and

$$(14) \quad C(u_{1t}, u_{2t}) = V(u_{1t})$$

Expression (14) implies, as noted above, that  $\beta = 1$ . The portfolio analysis of composite predictions then also leads to the conclusion that if the *FMP* model succeeds in utilizing the larger information set available to it, subsuming the information contained in *ARIMA* predictions, estimates  $\hat{\beta}_1$  and  $\hat{\beta}_2$  in (3) should be approximately unity and zero.

Least squares estimates of  $\beta_1$  and  $\beta_2$  in (3) for each of the endogenous variables appear in Table 3. Values of  $\hat{\beta}_1$  are significant at the 5 percent level for all of the variables. Values of  $\hat{\beta}_2$  are significant at the 5 percent level for nine of fourteen variables and at the 10 percent level for a tenth. Durbin-Watson statistics (*D-W*) are generally close to two, and in no case may the hypothesis that the errors of the composite prediction are uncorrelated be rejected at the 5 percent level.<sup>9</sup> These results suggest that the *ARIMA* predictions do embody information which is omitted by *FMP* predictions, in particular, information available from the history of the individual variables themselves.

Since individual predictions are essentially unbiased, we would expect that their coefficients in a composite would add to approximately unity. Tests of the hypothesis that  $\beta_1 + \beta_2 = 1$  in each regression led to rejection only in the case of non-farm inventory investment. The weights given to the *ARIMA* predictions when the weights are reestimated under the

<sup>9</sup> When constant terms were added to regressions (3) none were significantly different from zero.

TABLE 3—COMPOSITE SAMPLE PERIOD PREDICTIONS OF ENDOGENOUS VARIABLES

Endogenous Variables	Minimum Squared Error Composite Predictions				Weight Given to <i>ARIMA</i> Under Constraint $\hat{\beta}_1 + \hat{\beta}_2 = 1$
	Weights		Standard Deviation of Error	<i>D-W</i>	
	<i>FMP</i>	<i>ARIMA</i>			
1. <i>Gross National Product</i>	.834	.167	3.27	1.87	.168
2. Consumers' Expenditures on Nondurable Goods	.722	.278	1.36	2.39	.264
3. Consumers' Expenditures on Durable Goods	.958	.042	1.14	2.41	.044
4. Nonfarm Inventory Investment	.975	.194	2.38	1.81	.185
5. Expenditures on Producers' Durables	.658	.340*	.70	1.62	.369*
6. Expenditures on Producers' Structures	.670	.334*	.47	1.99	.355*
7. State and Local Government Expenditures	.290	.712*	.46	2.35	.681*
8. Housing Expenditures <sup>a</sup>	.794	.209*	.42	1.44	.225*
9. Unemployment Rate	.338	.662*	.27	1.91	.659*
10. <i>GNP</i> Deflator-Price Index	.641	.359*	.18	1.70	.363*
11. Consumer Goods Price Index	.561	.439*	.19	2.00	.436*
12. Yield on U.S. Treasury Bills	.706	.301*	.22	2.19	.292*
13. Yield on Commercial Paper	.736	.276*	.21	1.86	.273*
14. Yield on Corporate Bonds	.750	.255**	.09	1.85	.225

<sup>a</sup> Sample period for Housing Expenditures is 1957-01 through 1966-04.

\* Denotes weight for *ARIMA* prediction significant at the 5 percent level.

\*\* Denotes weight for *ARIMA* prediction significant at the 10 percent level.

constraint  $\hat{\beta}_1 + \hat{\beta}_2 = 1$  are also given in Table 3 and differ little from the unconstrained estimates.

#### IV. Jointly Optimal Composite Predictions

While the composite forecasts given by (3) are of interest in assessing the utilization of information by the *FMP* model, they may not be the optimal composites for a decision maker whose objective is to select weights which minimize expected loss. In particular, the relationships between errors for different variables may be of crucial importance as noted in Section II. A class of loss functions which allows for such interaction is that of the quadratic forms

$$(15) \quad L = \underline{\epsilon}' W \underline{\epsilon}$$

where  $L$  is the loss associated with the vector of errors across variables,  $\underline{\epsilon}$ , and  $W$

is a symmetric matrix. Enumeration of plausible choices of  $W$  is, of course, impossible. As an illustrative example, however, consider the particular loss function

$$(16) \quad L = \underline{\epsilon}' \Sigma^{-1} \underline{\epsilon}$$

where  $\Sigma$  is  $\text{Var}(\underline{\epsilon})$ , the matrix of variances and covariances of composite errors. Minimization of average loss over a given sample period corresponds in this special case to Aitken's generalized least squares estimation of parameters  $\beta_1$  and  $\beta_2$  for each of the equations (3) over the set of endogenous variables of interest. The matrix  $\Sigma$  is, of course, in practice unknown, and must be estimated. Zellner has suggested that  $\Sigma$  be estimated as the matrix of sample moments of residuals from ordinary least square estimation of the individual equations, in this case the individual composite predictions. Estimates of jointly

optimal weights obtained by Zellner's procedure appear in Table 4. The weight assigned to the *FMP* prediction is significant at the 5 percent level for each variable. The weight assigned to the *ARIMA* prediction is significant at the 5 percent level for ten of the fourteen variables and at the 10 percent level for another three. Thus, joint estimation of optimal weights for *ARIMA* predictions reinforces our conclusion that these predictions utilize information which is omitted by the *FMP* predictions.

The weights given in Table 4 sum to approximately unity for each variable except Nonfarm Inventory Investment. Individual *t*-statistics for the linear hypotheses  $\beta_1 + \beta_2 = 1$  are not significant except in the case of the latter variable. However, the *F*-statistic for the joint test of  $\beta_1 + \beta_2 = 1$  for all variables is 4.03 with 14 and 532 degrees of freedom so that we may reject the joint hypothesis at the .01

level. Thus, while departures from unbiasedness over the sample period may not be large in absolute magnitude, they are sufficient to provide rejection of the joint hypothesis of unbiasedness.

The general implication of stating the problem of composite weight selection in a loss function context is that from the viewpoint of the decision maker the question of whether one set of predictions or the other is more accurate is irrelevant. Since his objective is to minimize expected loss, he will purchase any piece of information which reduces expected loss by more than its cost. Thus, the value of the *ARIMA* predictions, for example, is not measured by their individual errors, but rather by the contribution which they are able to make to the reduction in expected loss associated with a composite prediction or a set of composite predictions. This is also true, of course, for the *FMP* predictions. Since the latter are rela-

TABLE 4—JOINTLY OPTIMAL COMPOSITE PREDICTIONS

Endogenous Variable	Weights for Jointly Optimal Composite Predictions <sup>a</sup>		<i>t</i> -statistic for Hypothesis $\beta_1 + \beta_2 = 1$
	<i>FMP</i>	<i>ARIMA</i>	
1. Gross National Product	.881	.119*	1.301
2. Consumers' Expenditures on Nondurable Goods	.807	.194**	.542
3. Consumers' Expenditures on Durable Goods	.936	.065	.056
4. Nonfarm Inventory Investment	1.042	.091**	4.424
5. Expenditures on Producers' Durables	.692	.306*	-.903
6. Expenditures on Producers' Structures	.659	.343*	.720
7. State and Local Government Expenditures	.345	.656	.931
8. Housing Expenditures	.880	.123**	1.263
9. Unemployment Rate	.310	.698*	1.052
10. GNP Deflator-Price Index	.791	.209*	.302
11. Consumer Goods Price Index	.711	.289*	.548
12. Yield on U.S. Treasury Bills	.668	.354*	.338
13. Yield on Commercial Paper	.700	.310*	1.121
14. Yield on Corporate Bonds	.816	.187*	.948

<sup>a</sup> Sample period for estimate of jointly optimal weights is 1957-01 through 1966-04.

\* Denotes weight for *ARIMA* prediction significant at the 5 percent level.

\*\* Denotes weight for *ARIMA* prediction significant at the 10 percent level.

tively expensive relative to *ARIMA* predictions (including computational expense, updating, etc.), we might expect to find that many decision makers would purchase the less accurate and less expensive set of predictions. Likewise, if the bum on the street corner offers free tips to the decision maker on his way to the office, these will be incorporated in composite predictions if they result in any reduction in expected loss, regardless of presumably gross inaccuracy.

### V. Analysis of Post-Sample Prediction Errors

It is scarcely surprising that both sets of predictors as well as their composites achieve reasonable accuracy during the period they were designed to explain. In the operational use of models, however, neither the forecaster nor the policy maker enjoys the luxury of working within the period of fit. Rather, from their point of view it is post-sample performance which is most relevant. Data for quarters 1967-01 through 1969-01 included in the *FMP* data deck provide only a short post-sample record, but nevertheless yield rather interesting and important results.

The mean squares, means, and standard deviations of post-sample one-quarter-ahead errors for both *FMP* and *ARIMA* models appear in Table 5 (*FMP* predictions continue to be conditioned on *true* future values of exogenous variables). It is immediately apparent that the accuracy of both sets of predictions deteriorated substantially during the post-sample period. However, mean square errors are smaller for *ARIMA* than for *FMP* predictions in the case of *GNP*, both categories of Consumer Expenditures, Expenditures on Producers' Durables, State and Local Government Expenditures, the Unemployment Rate, and all three interest rates. Differences are small for the *GNP* Deflator and Consumer Goods Price Index. It would appear, then, that the accuracy of *FMP* predictions deteriorated relative to that of *ARIMA* predictions during the post-sample period. In particular, the *FMP* model appears to have overestimated the effect of the federal tax surcharge enacted in 1968 and applied to personal income taxes in the third quarter of 1968 and to corporate income taxes retroactively to the first quarter of 1968. The *FMP* prediction of *GNP* was low by

TABLE 5—SUMMARY STATISTICS FOR *FMP* MODEL AND *ARIMA* MODEL POST-SAMPLE PREDICTION ERRORS

Endogenous Variables	<i>FMP</i> Model Errors			<i>ARIMA</i> Model Errors			Errors of Jointly Estimated Composite Predictions		
	<i>MSE</i>	Mean	Standard Deviation	<i>MSE</i>	Mean	Standard Deviation	<i>MSE</i>	Mean	Standard Deviation
1. Gross National Product	77.259	3.782	7.934	36.652	2.632	5.452	55.468	2.979	6.826
2. Consumers' Expenditures on Nondurable Goods	25.540	-3.914	3.197	11.605	1.464	3.076	18.944	-3.050	3.105
3. Consumers' Expenditures on Durable Goods	14.440	2.152	3.132	5.369	.990	2.095	13.270	2.065	3.001
4. Nonfarm Inventory Investment	11.161	1.471	2.998	49.589	-.166	7.040	12.285	.586	3.456
5. Expenditures on Producers' Durables	22.288	2.752	3.836	6.211	.668	2.401	15.939	2.261	3.291
6. Expenditures on Producers' Structures	1.038	.220	.995	4.427	.337	2.077	1.596	.349	1.214
7. State and Local Government Expenditures	8.065	.692	2.754	.766	.001	.875	1.414	.118	1.183
8. Housing Expenditures	1.935	1.002	.965	2.646	.764	1.436	1.572	.873	.899
9. Unemployment Rate	.412	-.522	.374	.081	-.141	.247	.114	-.287	.178
10. <i>GNP</i> Deflator-Price Index	.068	-.016	.260	.120	.237	.253	.051	.025	.225
11. Consumer Goods Price Index	.098	-.191	.249	.200	.295	.336	.068	-.074	.249
12. Yield on U.S. Treasury Bills	.425	.176	.628	.305	.091	.545	.280	.132	.512
13. Yield on Commercial Paper	.240	.282	.400	.190	.085	.427	.168	.172	.372
14. Yield on Corporate Bonds	.066	.150	.204	.055	.116	.205	.058	.133	.200

\$4.9 billion and \$4.4 billion in the third and fourth quarters of 1968 compared with \$5.5 billion and \$2.4 billion, respectively, for the *ARIMA* model. In the first quarter of 1969 the *FMP* model got very seriously off-track with a prediction that was too low by \$23.2 billion when the *ARIMA* prediction was high by \$1.9 billion.

The results described above suggest that the simple *ARIMA* models are relatively more robust with respect to post-sample prediction than is the complex *FMP* model. It is interesting that this comparison generalizes to a considerable extent to relative performance among *ARIMA* models for different variables. In particular, the ratios of post-sample to sample Mean Square Error (*MSE*) are large for the fairly complex models for Expenditures on Producers' Structures and Housing Expenditures. Among the best performers are the very simple models for *GNP* and Consumers' Expenditures on Durable Goods, although the four-parameter model for the Unemployment Rate is the best of all and is the only model with a post-sample *MSE* smaller than its sample period *MSE*.

Finally, it is interesting to compare the post-sample performance of *FMP* and *ARIMA* predictions with that of the composite predictions constructed using the jointly estimated weights of Table 4. The relative magnitudes of mean square errors given in Table 5 indicate that composite predictions were more accurate than *FMP* predictions for twelve of the fourteen variables, the exceptions being Nonfarm Inventory Investment and Housing Expenditures for which the *ARIMA* component had suffered considerable post-sample deterioration. Composite predictions were more accurate than *ARIMA* predictions, however, in only seven cases, reflecting the generally severe deterioration in *FMP* performance. Composite predictions were more accurate than

either individual prediction in five cases. If we score each of the three predictions by number of first places, the *ARIMA* models earn seven points, the composites five, and the *FMP* model only two. Thus, if mean square error were an appropriate measure of loss, an unweighted assessment clearly indicates that a decision maker would have been best off relying simply on *ARIMA* predictions in the post-sample period. To have ignored the information available from the simple time series models altogether would have been costly indeed.

#### APPENDIX

The following are the *ARIMA* models fitted on data from the *FMP* data deck for 1947-01 through 1966-04. Variables (1) through (8) are in billions of current dollars, the remaining variables in percentage points. The  $z_t$  and  $a_t$  are understood to be general notation referring to the observed value and disturbance associated with each respective variable.

##### 1. Gross National Product

$$z_t = z_{t-1} + .615(z_{t-1} - z_{t-2}) \\ + 2.76 + a_t$$

$$\hat{\sigma}_a = 4.77$$

##### 2. Consumers' Expenditures on Nondurable Goods

$$z_t = z_{t-1} + .190(z_{t-1} - z_{t-2}) \\ + .504(z_{t-2} - z_{t-3}) + 1.06 + a_t$$

$$\hat{\sigma}_a = 1.72$$

##### 3. Consumers' Expenditures on Durable Goods

$$z_t = z_{t-1} + .666 + a_t$$

$$\hat{\sigma}_a = 1.92$$

##### 4. Nonfarm Inventory Investment

$$z_t = .581z_{t-1} + a_t + .0013a_{t-1} \\ + .742a_{t-2} + 1.69$$

$$\hat{\sigma}_a = 3.14$$

## 5. Expenditures on Producers' Durables

$$z_t = z_{t-1} + a_t + .347a_{t-1} + .517$$

$$\hat{\sigma}_a = 1.06$$

## 6. Expenditures on Producers' Structures

$$z_t = z_{t-1} + .303(z_{t-1} - z_{t-2})$$

$$+ .216(z_{t-2} - z_{t-3})$$

$$+ .297(z_{t-3} - z_{t-4})$$

$$- .442(z_{t-4} - z_{t-5})$$

$$+ .159 + a_t$$

$$\hat{\sigma}_a = .47$$

## 7. State and Local Government Expenditures

$$z_t = 2z_{t-1} - z_{t-2} + a_t - .695a_{t-1}$$

$$\hat{\sigma}_a = .52$$

## 8. Housing Expenditures

$$z_t = z_{t-1} + .639(z_{t-1} - z_{t-2})$$

$$+ .076(z_{t-2} - z_{t-3})$$

$$- .286(z_{t-3} - z_{t-4}) + a_t$$

$$\hat{\sigma}_a = .74$$

## 9. Unemployment Rate

$$z_t = 1.46z_{t-1} - .612z_{t-2} + a_t$$

$$+ .284a_{t-1} + .734$$

$$\hat{\sigma}_a = .33$$

## 10. GNP Deflator-Price Index

$$z_t = z_{t-1} + .523(z_{t-1} - z_{t-2}) + a_t + .256$$

$$\hat{\sigma}_a = .46$$

## 11. Consumer Goods Price Index

$$z_t = z_{t-1} + .414(z_{t-1} - z_{t-2}) + a_t + .244$$

$$\hat{\sigma}_a = .48$$

## 12. Yield on U.S. Treasury Bills

$$z_t = z_{t-1} + .608(z_{t-1} - z_{t-2})$$

$$- .425(z_{t-2} - z_{t-3}) + a_t$$

$$\hat{\sigma}_a = .29$$

## 13. Yield on Commercial Paper

$$z_t = z_{t-1} + .727(z_{t-1} - z_{t-2})$$

$$- .427(z_{t-2} - z_{t-3}) + a_t$$

$$\hat{\sigma}_a = .27$$

## 14. Yield on Corporate Bonds

$$z_t = z_{t-1} + .490(z_{t-1} - z_{t-2})$$

$$- .169(z_{t-2} - z_{t-3}) + a_t$$

$$\hat{\sigma}_a = .11$$

## REFERENCES

- G. E. P. Box and G. M. Jenkins, *Time Series Analysis, Forecasting and Control*, San Francisco 1970.
- J. P. Cooper and S. Fischer, "Stochastic Simulation of Monetary Rules in Two Macroeconometric Models," Univ. of Chicago Center for Mathematical Studies in Business and Economics, Rep. 7106, Chicago 1971.
- F. DeLeeuw and E. M. Gramlich, "The Federal Reserve—MIT Econometric Model," *Fed. Reserve Bull.*, Jan. 1964, 54, 11-40.
- R. Frisch, "Propagation Problems and Impulse Problems in Dynamic Economics," in *Economic Essays in Honor of Gustav Cassel*, London 1933, 171-205.
- D. M. Grether and M. Nerlove, "Some Properties of 'Optimal' Seasonal Adjustment," *Econometrica*, Sept. 1970, 38, 682-703.
- F. Modigliani, R. Rasche, and J. P. Cooper, "Central Bank Policy, the Money Supply and the Short-Term Rate of Interest," *J. Money, Credit Banking*, May 1970, 2, 116-218.
- G. H. Moore, "Forecasting Short-Term Economic Change," *J. Amer. Statist. Ass.*, Mar. 1969, 64, 1-22.
- J. P. Muth, "Rational Expectations and the Theory of Price Movements," *Econometrica*, July 1961, 29, 315-35.
- C. R. Nelson, *The Term Structure of Interest Rates*, New York 1972.
- R. H. Rasche and H. T. Shapiro, "The FRB-MIT Econometric Model: Its Special Features," *Amer. Econ. Rev. Proc.*, May 1968, 58, 123-49.

- E. E. Slutsky, "The Summation of Random Causes as the Source of Cyclical Processes," *Problems of Economic Conditions*, The Conjecture Institute, 3, No. 1, Moskva (Moscow) 1927; reprinted in *Econometrica*, Apr. 1937, 5, 105-146.
- H. O. Steckler, "Forecasting with Econometric Models: An Evaluation," *Econometrica*, July-Oct. 1968, 36, 437-63.
- H. Theil, *Applied Economic Forecasting*, Amsterdam 1966.
- V. Zarnowitz, *An Appeal of Short-Term Economic Forecasts*, New York 1967.
- A. Zellner, "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias," *J. Amer. Statist. Ass.*, June 1962, 57, 348-68.
- , "General Description of the Federal Reserve-MIT-Penn Quarterly Econometric Model of the U.S. Economy (Version 4.1-4/15/69)," mimeo working paper, Univ. Chicago 1969.
- and S. Peck, "Simulation Experiments with a Quarterly Macro-Econometric Model of the U.S. Economy," mimeo working paper, Univ. Chicago 1971.

# Uncertain Entry and Excess Capacity

By MORTON I. KAMIEN AND NANCY L. SCHWARTZ\*

Despite shifts in the focus of economic research, there are some problems that successive generations of economists address, among them the "excess capacity" proposition enunciated by Edward Chamberlin. Since its introduction, leading theorists have debated, modified, and clarified this proposition. Syntheses by William J. Baumol (1964) and Jagdish Bhagwati have appeared recently.

In brief, the excess capacity proposition is an assertion that a firm in monopolistic or imperfect competition will operate a productive facility at a lower rate than that at which average cost achieves its minimum. It is argued that the presence of economic profit attracts entry of rivals until that profit is eliminated. But zero profit can be attained by a profit-maximizing firm only where average cost and demand curves are tangent. Since the demand curve must be falling at that tangency point, so must the average cost curve.

Harold Demsetz has attempted a refutation of the excess capacity proposition. He contends that the relevant cost function includes promotional costs as well as production costs, and claims the excess capacity proposition cannot be inferred when the extended cost function is employed. This argument has been challenged recently by Yoram Barzel.

Another reappraisal of the proposition's applicability originates with Nicholas Kalder and has been expanded successively by

Roy Harrod, J. R. Hicks, and Frank Hahn. They argue that since existing firms are differentiated, any particular firm is faced with close and distant rivals in terms of similarity of product, location, or custom. Likewise, entry of a new rival typically would not have equal (negligible) impact on all existing firms, but instead would significantly affect one or several firms already in the industry with less impact on the remainder. Hence, it is contended, an existing firm would recognize the possibility of close rival entry and would attempt to forestall such entry and/or take that possibility into account when selecting plant size. Because the pre- and post-entry output rates will differ, plant size cannot be optimal for both periods, and in fact will not be optimally chosen for either output rate taken by itself. This argument does lead to the conclusion that post-entry capacity may be excessive, although not for the Chamberlinian reason outlined at the beginning of this paper.

In this paper we explicitly allow for the possibility that the firm does not know when rivals will appear. We also assume that the firm can retard rival appearance through its pricing policy; the probability of entry at any time is supposed directly related to the price the firm sets. The firm is to select both a pre-entry price and a capital stock to maximize the present value of its expected cash flow over the indefinite future. In analyzing this model, we will indicate the sense in which there can be excess capacity in either the pre-entry or post-entry period. The dependence of the optimal capacity upon the production plan in both pre- and post-

\* Professors of managerial economics, Graduate School of Management, Northwestern University. We gratefully acknowledge research support from the National Science Foundation. Responsibility for opinions expressed remains with the authors. We also appreciate the referee's suggestions.

entry periods, ease of entry, the discount rate, and uncertainty regarding the time of rival entry will be displayed. The model explains why firms fail to maximize current profits. It also shows that a firm will price according to a demand schedule more elastic than the market demand for its product. Finally we indicate that our principal conclusions about capacity obtain if the firm attempts to retard entry by means other than pricing.

### I. The Model

The argument of Hicks and Hahn mentioned above constitutes the starting point of our investigation. We posit a firm operating in an imperfectly competitive environment, by which we mean the firm has actual and/or potential rivals so that perfect competition does not prevail. For definiteness, we suppose the firm currently holds a monopoly position in some market or submarket, or equivalently, that there are several firms which tacitly collude to maximize their long-run profits within that submarket. In addition, it is supposed that the possibility of rival entrants exists. The existing firm views the probability of rival entry as an increasing function of the price it sets. It will be shown later how to modify the analysis to accommodate the manipulation of some parameter other than price, such as advertising, to retard the appearance of rivals.

Our selection of price rather than profit as the signal to potential entrants is based on both its greater visibility and the argument that the potential entrant is interested in the profits he can make, not the current profits realized by existing firms. The cost of production of a new firm could differ from that of the existing firm. Thus entry into an industry in which profits are not high is possible if the entrant could produce at a cost which provides him a profit at a price below that

currently prevailing. If the current firm's costs are known by others, "price" and "profits" of course convey the same information. In addition, David Baron has observed that the supposition that profit attracts entry leads to peculiar conclusions regarding pricing. If profit is a concave function of price with a finite maximum and if the probability of entry is an increasing function of profit, the firm can obtain an entry-retarding profit equally well by setting price below or above the current profit maximizing level.

While price may be freely varied, plant size cannot be altered at all, once selected. It is supposed that the plant is infinitely durable and will be used over the indefinite future without modification. (Thus we do not consider the case in which it is preferable to scrap the old plant and build a new one after rivals have entered.) Varying amounts of current inputs may be employed with the fixed plant to produce a range of output rates. The assumption that plant size is fixed once it has been selected is admittedly a polar case, consistent with previous investigations. As this supposition is relaxed, the tendency towards excess capacity, as defined in this paper, declines and with perfect flexibility in plant adjustment, disappears altogether; see Graham Pyatt.

The future is divided into two periods, with the interval (of unknown length) from time zero until the appearance of an entrant called period 1, and the remaining time span denoted period 2. The firm's choice of output rates prior to and after entry of rivals will not in general be identical. There seems to be no consensus about the relative magnitudes of these rates. It has sometimes been supposed that the firm's output rate would decline when the monopoly was lost since the market would then be shared among more sellers, but others believe output will rise since the entrants get only part of the marginal

demand (demand at prices below that prevailing in the pre-entry period), leaving the remainder for the original firm. It will not be necessary for us to take a position on this question.

To express our model more precisely, let  $k$  denote the amount of capital to be employed over the future. (This variable also serves as an index of size of plant, where size may be measured by the output level at which average cost of production with the given plant is minimized.) Let  $q^n$ ,  $p^n$ ,  $Q^n(p^n)$  denote the rate of output, product price, and the product demand function respectively in the  $n$ th period, where  $n=1, 2$ . The demand function and the corresponding marginal revenue function are supposed negatively related to price, and continuously differentiable. The product demand function is assumed stationary, but output and price could vary within periods. There are  $m$  variable inputs which may be purchased in any quantities  $x=(x_1, \dots, x_m)$ , at constant prices  $w=(w_1, \dots, w_m)$ . The constant purchase price of a unit of capital is denoted  $W$ , while  $r$  is the (constant) rate at which future cash flows are discounted. The rate of output is related to the utilization of inputs through the production function  $f(k, x)=q$ . The production function is supposed twice differentiable and concave. Demand must be satisfied by current production, ruling out storage.

We assume that the probability of rival entry at time  $t$ , given that entrants have not previously appeared, is an increasing, convex function  $h(p^1)$  of product price  $p^1(t)$  at time  $t$ . Letting  $F(t)$  denote the probability that rival entry will have occurred by date  $t$ , the conditional entry probability is  $F'(t)/(1-F(t))$ , where prime indicates differentiation. Thus

$$F'(t)/(1-F(t)) = h(p^1(t))$$

where  $h(0)=0$ ,  $h'(p^1) \geq 0$ ,  $h''(p^1) \geq 0$  for  $p^1 \geq 0$ .

Since the exact time of entry is not known, the firm must develop a policy applicable while the monopoly is maintained and a policy for after the entrant appears. Cash flow at time  $t > 0$  will be  $Q^1 p^1 - wx^1$  if entry has not yet occurred and  $Q^2 p^2 - wx^2$  if entry has occurred. The firm's problem is to select a single capital stock  $k$ , variable input rates  $x^1(t)$ ,  $x^2(t)$ , and price policies  $p^1(t)$ ,  $p^2(t)$  to maximize the present value of its expected cash flow over the entire future, subject to the relationship between  $p^1(t)$  and rival entry and to the requirement that production meet demand. The problem may be posed formally as

- (1)  $\max -Wk$   
 $+ \int_0^\infty e^{-rt} \{ [Q^1(p^1(t))p^1(t) - wx^1(t)](1-F(t))$   
 $+ [Q^2(p^2(t))p^2(t) - wx^2(t)]F(t) \} dt$
- (2) subject to  $F'(t) = h(p^1(t))(1-F(t))$
- (3)  $F(0) = 0$
- (4)  $Q^n(p^n(t)) = f(k, x^n(t)) \quad n=1, 2$

The above formulation represents an optimal control problem with a single state variable  $F(t)$ . Equation (3) reflects the assumption that at the initial time, rival entry has not occurred. It is possible to modify the formulation slightly to allow for a time span that precedes the onset of production, as for example due to time required to construct the plant. This is accomplished by rewriting (3) as  $F(0) = F_0$ , where  $F_0$  is the probability of rival entry during the period between planning and the beginning of production.

The investment required initially to obtain the  $k$  units of capital (the plant of size  $k$ ) is  $Wk$ . If that investment were to be financed over time, then reinterpret  $W$  as the present value of the outlays required to obtain a unit of capital at time 0.

## II. Solution

Rather than present the necessary conditions corresponding to problem (1)–(4) immediately, we proceed stepwise, noting successive simplifications in the problem statement. Ultimately, necessary conditions are provided through study of stepwise interrelated problems in a finite number of variables.

It is evident on inspection of (1)–(4) that current factors  $x^1(t)$ ,  $x^2(t)$  are selected to minimize the variable costs of producing at the rates  $q^n(t) = Q^n(p^n(t))$ ,  $n = 1, 2$ , respectively, with the existing capital stock  $k$ . Denote the minimum variable cost of producing at rate  $q$  with plant size  $k$  as

$$(5) \quad V(q; k) = \min_x wx$$

$$\text{subject to } f(k, x) = q$$

For given factor prices the minimizing input combination determined in (5) will depend on  $q$  and  $k$ , i.e.,  $x = X(q, k)$ . The rate of cash flow or quasi rent is a function of product price and size of plant. Denote this cash flow as

$$(6) \quad \pi^n(p^n, k) = Q^n(p^n)p^n - V(Q^n(p^n); k) \\ n = 1, 2$$

The foregoing observations and definitions enable us to rewrite the problem (1)–(4) as one of selecting a capital stock  $k$  and price policies  $p^1(t)$ ,  $p^2(t)$  to

$$(7) \quad \max -Wk + \int_0^\infty e^{-rt} [\pi^1(p^1(t), k)(1 - F(t)) \\ + \pi^2(p^2(t), k)F(t)] dt$$

$$\text{subject to } F'(t) = h(p^1(t))(1 - F(t))$$

$$F(0) = 0$$

The variables  $x^n(t) = X(Q^n(p^n(t)), k)$ ,  $n = 1, 2$  are completely determined from (5) and the solution to (7).

Once entry has occurred, there is no reason for the firm to choose a price below that which will maximize current cash flow  $\pi^2$ . Further, since that function does not depend explicitly on time, the post-entry price policy will be constant with respect to  $t$ ,  $p^2(t) = p^2$ . See (7).

To the suppositions already made, one is now added which enables us to conclude that under all our assumptions, the optimal price policy is stationary in period 1 as well as in period 2. In particular, we suppose that the equation

$$\pi^1(p^1, k) = b,$$

where

$$0 \leq b \leq \max_{p^2} \pi^2(p^2, k),$$

has two distinct positive solutions  $p^1$  for all  $k > 0$ . This assumption also reflects our hypothesis that maintenance of a monopoly position can be worthwhile; that is, it is possible to make a higher return before the appearance of rivals than will be attainable afterwards. It can then be shown, following the method of proof employed in Kamien and Schwartz, that the optimal pre-entry price policy is a constant function of time.

Since  $p^1(t) = p^1$ , the conditional probability of rival entry is likewise a constant function of time,  $h(p^1)$ . Solving (2)–(3) with  $p^1$  constant, the probability distribution governing rival entry is

$$(8) \quad F(t) = 1 - e^{-h(p^1)t}$$

On substituting (8) into the objective function (7), and integrating, we obtain expression for the present value of the expected profits stream

$$(9) \quad -Wk + \frac{\pi^1(p^1, k) - \pi^2(p^2, k)}{r + h(p^1)} \\ + \frac{\pi^2(p^2, k)}{r}$$

which is to be maximized by choice of pre-

entry and post-entry prices  $p^1$ ,  $p^2$ , and capital stock  $k$ . The stream of net receipts or quasi rents is composed of the flow  $\pi^2$  which will be received indefinitely plus an extra amount  $\pi^1 - \pi^2$  which will be collected only so long as the monopoly is maintained. The former quantity is a certain return, and is capitalized at the discount rate  $r$ . The temporary or transient addition has expected duration  $1/h(p^1)$  and is capitalized at a higher rate  $r+h(p^1)$ .

Thus we have simplified the original problem statement (1)–(4) to the maximization of (9). To find necessary conditions for a maximum of (9), we substitute from definition (6) and differentiate the resulting expression with respect to  $p^1$ ,  $p^2$ , and  $k$ . Setting those derivatives equal to zero yields, respectively,

$$(10) \quad Q_p^1[p^1 + Q^1/Q_p^1 - V_q^1] \\ = [\pi^1 - \pi^2]h_p/(r+h)$$

$$(11) \quad p^2 + Q^2/Q_p^2 = V_q^2$$

$$(12) \quad -(rV_k^1 + hV_k^2)/(r+h) = rW$$

where subscripts indicate differentiation, arguments have been suppressed, and the superscript  $n$  on  $V$  indicates that it is to be evaluated at  $q = Q^n(p^n)$ .

We can relate these conditions to the more elemental components of the model, the production function and factor prices, by observing that if the minimizing input vector  $x^n$  is strictly positive as we assume, then from (5) it must satisfy

$$w = \lambda^n f_x(k, x^n)$$

and the constraint

$$f(k, x^n) = q^n = Q^n(p^n)$$

Here  $f_x = (\partial f/\partial x_1, \dots, \partial f/\partial x_m)$  and  $\lambda^n$  is the Lagrange multiplier associated with the constraint. Moreover,  $\lambda^n = V_q^n$  since we may write

$$(13) \quad V(q; k) = wx + \lambda[q - f(k, x)]$$

where  $x$  is optimal. Differentiate with respect to  $q$  to obtain

$$V_q(q; k) = wx_q + \lambda_q[q - f(k, x)] \\ + \lambda[1 - f_x(k, x)x_q]$$

where  $x_q = (\partial x_1/\partial q, \dots, \partial x_m/\partial q)$ . Since  $x$  is optimal, the coefficients of  $x_q$  and  $\lambda_q$  both vanish. Hence the above equation reduces to  $V_q = \lambda$  as claimed. Consequently

$$(14) \quad w = V_q^n f_x(k, x^n) \quad n = 1, 2$$

In similar fashion, (13) may be differentiated with respect to  $k$ , to see how the minimum variable cost of producing at rate  $q$  changes with the capital stock:

$$V_k(q; k) = wx_k + \lambda_k[q - f(k, x)] \\ - \lambda[f_k + f_x(k, x)x_k]$$

where subscripts indicate differentiation as above. Since  $x$  is optimal, the coefficient of  $x_k$  and the coefficient of  $\lambda_k$  equal zero so that

$$(15) \quad V_k^n = -V_q^n f_k(k, x^n) \quad n = 1, 2$$

We have shown that  $V_q$  and  $V_k$  may be expressed in terms of factor prices and marginal productivities. Thus, we can completely summarize our necessary conditions as follows: If there is a positive finite solution to the problem (1)–(4), the optimal values of the variables  $p^1$ ,  $p^2$ ,  $x^1$ ,  $x^2$ ,  $k$  will be constant functions of time, which together with variables  $V_k^n$ ,  $V_q^n$ ,  $n=1, 2$  satisfy equations (4), (10)–(12), (14)–(15).

### III. Discussion

We now proceed to interpret and discuss the implications of the necessary conditions (4), (10)–(12), (14)–(15). Equations (4) have been discussed previously. Conditions (14) are the usual requirement that the marginal physical product of each variable factor be proportional to its price. The common factor of proportionality,

$V_a$ , is the (short-run) marginal cost of production. Likewise (11) expresses the condition that marginal revenue equal marginal cost in the second period. Quasi rent is maximized in period 2 since, after a rival appears, nothing is gained by foregoing current return. It follows from (14) that  $V_k^n > 0$  and hence from (15) that  $V_k^n < 0$  for  $n=1, 2$ . According to (15), the reduction in the variable cost of producing at any given rate due to slight expansion of capital is equal to the marginal product of capital multiplied by (short-run) marginal cost.

The two conditions that differ from typical neoclassical requirements are (10) and (12). First period price policy affects both the magnitude  $\pi^1 - \pi^2$  of the transient return and its expected duration  $1/h(p^1)$ . Thus, equation (10) requires that pre-entry price cause the proportionate rate of increase in the capitalization rate of the transient return due to a price increase to equal the proportionate increase in the magnitude of the transient return due to that price increase, i.e.,

$$d \ln(r + h(p^1))/dp^1 = d \ln(\pi^1 - \pi^2)/dp^1$$

If  $h_p > 0$  and  $\pi_1 > \pi_2$ , then both sides of (10) will be positive and marginal cost will exceed marginal revenue in the first period. This means that price will be lower and output higher in period 1 than if current quasi rent to the same capital stock were being minimized. If, alternatively,  $h_p = 0$ , which reflects the existence of natural barriers to entry that assure a monopoly, then marginal cost will be equated with marginal revenue in period 1. In addition,  $h_p = 0$  implies  $h = 0$  which means that the firm will also pick a different capital stock than in the previous case, see (12). Consequently we cannot compare first period price under the alternative circumstances. See G. C. Archibald. We also note that (10) and (13) with  $h_p = h = 0$ , are identical to the first-order conditions of a firm that

completely ignores the possibility of rival entry. Equation (10) has been discussed more fully for the case of given capital stock in Kamien and Schwartz.

The major conclusion of our analysis regarding excess capacity stems from condition (12). According to (12), capital stock is to be selected so that a convex combination of the marginal cost saving of an incremental unit of capital in the two periods equals the average cost of employing a unit of capital. This implies

$$-V_k^1 \geq rW \geq -V_k^2$$

or

$$-V_k^1 \leq rW \leq -V_k^2$$

so that  $V_k^1 + rW$  and  $V_k^2 + rW$  are of opposite sign. Since these expressions may be interpreted as the partial derivative of total cost (of producing at rate  $q$  with capital  $k$ ) with respect to capital, we have obtained Hahn's conclusion (pp. 232-33). Additional insights are available from our conditions however. Using (15) to eliminate  $V_k^n$  from (12) we obtain

$$(16) \quad \frac{rV_{qfk}^1(k, x^1) + hV_{qfk}^2(k, x^2)}{r + h} = rW$$

The left side of this equation is a convex combination of "marginal cost times marginal product of capital" in each of the two periods with  $r/(r+h)$  and  $h/(r+h)$  the respective weights. If the firm is certain that no entry will occur,  $h(p^1) = 0$ , then (16) reduces to

$$(17) \quad V_{qfk}^1(k, x^1) = rW,$$

a condition completely analogous to the marginal productivity condition (14) for the variable factors. Since (17) holds when capacity is just suited to period 1 output and that output rate is expected to last forever, then by virtue of the diminishing

marginal productivity of capital assumption,  $f_{kk} < 0$ , we can take  $f_k > rW/V_q$  to mean that capacity is inadequate for the current output rate and the reverse inequality to mean there is excess capacity. It follows that

$$(18) \quad V_{qf_k}^1(k, x^1) \geq rW \geq V_{qf_k}^2(k, x^2) \\ \text{as } Q^1(p^1) \geq Q^2(p^2)$$

Thus unless output is identical in the two periods, there may be excess capacity in one of the periods and inadequate capacity in the other.

The average flow cost of using a unit of capital is set equal to a weighted average of  $V_{qf_k}^n(k, x^n)$ ,  $n=1, 2$ . The weights applied to the two periods depend on both the interest rate and on the expected length of the first period. The larger the interest rate and the more remote the expected loss of the monopoly, the larger the relative weight applied to the marginal productivity of capital and short-run marginal cost in the first period. In this sense, we can say that the more heavily the future is discounted and the longer the monopoly is expected to be maintained, the more the plant selected will conform to period 1 needs. Likewise, the smaller the interest rate and the sooner rivals are expected, the more closely the plant will suit period 2 output. Of course, the amount of excess capacity will depend also upon the production function and factor prices.

We are also able to separate the effect of pricing to retard entry, uncertainty, and multiperiod planning upon our results.

First, suppose the conditional entry probability  $h$  were a given positive constant, independent of price  $p^1$ . In that case,  $h_p = 0$  so that (10) reduces to the same form as (11); quasi rents will be maximized in period 1 and period 2. The form of the other conditions remains unchanged, although the values of the vari-

ables will be affected since those values are determined through simultaneous solution of all the necessary conditions. In particular, the form of (16), which may be thought of as the capital stock equation, is unchanged.

Second, to distinguish between those features of our results attributable to multiperiod planning and those caused by uncertainty regarding rival entry, we rewrite (16) in the form

$$(19) \quad \int_0^\infty h(p^1) e^{-h(p^1)T} \left[ \int_0^T e^{-rt} V_{qf_k}^1(k, x^1) dt \right. \\ \left. + \int_T^\infty e^{-rt} V_{qf_k}^2(k, x^2) dt \right] dT = W$$

According to (19) capital is to be purchased until its unit price  $W$  equals a discounted lifetime sum of its expected marginal productivity multiplied by the corresponding short-run marginal cost. The bracketed quantity is the relevant sum, if entry occurs at  $T$ , while the outer expression contains the probability density of  $T$  for a specified  $p^1$ . Suppose now that rival entry at a given date is certain. From the previous paragraph it is clear that in this event the firm will set price to maximize cash flow in each period. Capital stock will be selected by equating the bracketed term in (19) with  $W$ . Both results can be verified formally by recalling the problem statement in (7) and observing that  $F(t) = 0$  for  $t < T$ , and  $F(t) = 1$  for  $t \geq T$ . The other first-order conditions will be identical to those obtained under the uncertainty formulation.

The essential consequence of multiperiod planning alone is that capital is to be purchased until a weighted sum of its marginal productivities equals its price. This condition is analogous to the first-order condition for the production of a public good (see Paul Samuelson). The characteristic of a public good that gives rise to equation of a sum of individual

marginal rates of substitution with a marginal rate of transformation is that its consumption by one individual does not diminish its availability to another. Likewise, our assumption of infinite durability of capital implies that its use in one period does not deplete the amount available in a subsequent period. It is evident from (19) that introduction of uncertainty regarding rival entry into the model serves to modify the weights of the marginal productivities of capital summed.

#### IV. On Alternative Views and Extensions

The model formulated and discussed above may be used to provide an alternative explanation for certain phenomena discussed in the economics literature. In this section, we shall address two such areas and indicate an extension of our model to incorporate nonprice impediments to entry.

In the literature on monopolistic competition, it is typically assumed that the individual firm acts as though it faces a more elastic demand than in fact it does confront. The explanation given is that the firm imagines that it has no close rivals and so its actions and those of other industry members are unrelated. The original source of this explanation appears to be Chamberlin. We can provide an alternative explanation for this apparent behavior. Suppose a firm selects price  $p^1$  and output  $Q^1(p^1) = q^1$ , along with other variables according to conditions (4), (10)–(12), (14)–(15). The elasticity of the true demand function  $Q^1$  at this point we denote as

$$E^a = -p^1 Q_p^1(p^1)/Q^1(p^1)$$

On the other hand, if one were to assume that a firm producing at rate  $q^1$ , selling at price  $p^1$ , and incurring marginal cost  $V_q^1$  were maximizing current profits, then one would infer a different elasticity. In par-

ticular, the familiar "marginal revenue equals marginal cost" rule yields

$$p^1(1 - 1/E^s) = V_q^1$$

where  $E^s$  denotes the price elasticity of the "surrogate" demand which the firm is imagined to regard as relevant. Rearranging this marginal equality gives expression for the surrogate elasticity inferred as

$$E^s = p^1/(p^1 - V_q^1)$$

The quantity in square brackets on the left side of (10) is known to be negative; it may be rewritten as

$$\begin{aligned} 0 &> p^1(1 - 1/E^s) - V_q^1 \\ &= p^1[(p^1 - V_q^1)/p^1 - 1/E^s] \\ &= p^1(1/E^s - 1/E^a) \end{aligned}$$

from which it follows that  $E^s > E^a$  as claimed. Thus the firm which maximizes present value of long-run profits by taking potential rivals into account will appear to be trying to maximize current profits with a more elastic demand than it does face in fact. This conclusion is consistent with Kaldor's.

Next, we note that some economists maintain that firms do not maximize current profits, because they produce a larger quantity than is optimal under instantaneous profit maximization (with the given plant). Baumol (1967) has proposed a model based upon a sales maximization objective to explain this phenomenon. Our model of pricing to retard rival entry is an alternative explanation for selection of a price below that which maximizes current profit. In contrast to the sales maximization approach, our hypothesis does not deny the validity of profit maximization but rather takes a long-run view.

Finally, we note that the results obtained above regarding the firm's choice of

$$(20) \quad \frac{[Q^1(p^1)p^1 - w(x^1 + y)] - [Q^2(p^2)p^2 - wx^2]}{r + H(g(y))} + \frac{Q^2(p^2)p^2 - wx^2}{r} - Wk$$

$$(21) \quad \frac{w_i}{[Q^1(p^1)p^1 - w(x^1 + y)] - [Q^2(p^2)p^2 - wx^2]} = \frac{-H_u g_i}{r + H(g(y))} \quad \text{if } y_i > 0$$

plant scale can be derived under the alternative assumption that it postpones rival entry by erecting nonprice impediments. Such impediments may be created through advertising and promotional efforts which serve to build "brand loyalty" in consumers. Let  $u$  be a parameter reflecting the degree of exclusion or difficulty a potential entrant would have and suppose that the conditional entry probability  $H(u)$  at any time is a decreasing function of the current value of the exclusion parameter  $u$ . Thus the larger the parameter  $u$ , the smaller the conditional probability of rival entrants at that time. Let  $g(y) = u$  be the function relating the amounts of variable factors  $y = (y_1, \dots, y_m)$  devoted to exclusion activities to the degree of exclusion  $u$  achieved. The functions  $H(u)$  and  $g(y)$  are supposed concave in their arguments. The present value of the expected stream of cash flows is then, analogous to (9), given by equation (20) which is to be maximized through choice of  $p^n$ ,  $x^n$ ,  $y$ ,  $k$  ( $n=1, 2$ ) subject to (4). Forming the Lagrangian and setting its partial derivatives equal to zero yields the usual conditions regarding employment of current factors, as in (14), and that marginal revenue equal marginal cost in each period, as in (11). There is an additional stipulation that each resource  $y_i$  devoted to entry-retarding activities be employed up to the point at which the proportion by which it reduces transient cash flow is just equal to the proportionate rate of change in the rate of decay of that transient return, as seen in equation (21), where  $H_u = dH(u)/du$  and  $g_i = \partial g(y)/\partial y_i$ . Again scale will be intermediate between that op-

timal for first-period output and that for second-period output, with the weights applied to each period depending on the hazard rate and discount rate in the same fashion as in (16):

$$\frac{r\mu^1 f_k(k, x^1) + H(g(y))\mu^2 f_k(k, x^2)}{r + H(g(y))} = rW$$

where  $\mu^n$  is the marginal cost in period  $n$ . A model in which both price and nonprice activities are employed to retard rival entry can be constructed in an obvious manner. Since this model yields no additional insights, we have not presented it.

## V. Summary

The question of optimal plant size when that plant is to be used to produce at several output rates has been considered. We have studied a firm which recognizes the possibility that rivals will enter its market, but is uncertain about the entry date. Pricing policy does affect the probability of rival entry. The firm selects both a pre-entry price and a capital stock to maximize the present value of its expected cash flow over the indefinite future. The consequences of price upon the probability of rival entry and the implications of rival entry for appropriate plant scale are taken into account.

The optimal pre-entry price will be such that marginal cost exceeds marginal revenue. Hence price is lower and output higher than if the current return to the same capital stock were being maximized. We thus have obtained an alternative hypothesis to the sales maximization theory of Baumol to explain failure to maximize instantaneous profits. In addi-

tion we have an alternative hypothesis to the Chamberlinian  $dd'$ ,  $DD'$  demand curves to explain pricing according to a demand schedule which is more elastic than the demand schedule actually faced.

The capital stock is chosen so that a weighted sum of capital's marginal productivity is equal to the purchase price of a unit of capital. The weights depend on both the discount rate and the conditional entry probability at any time (hazard rate). In particular, the more heavily the future is discounted and the longer the monopoly is expected to be maintained, the more the plant selected will conform to pre-entry needs. Likewise, the smaller the interest rate and the sooner rivals are expected, the more closely the plant will suit post-entry output. Our conclusions are consistent with those obtained by Hahn with a model in which uncertainty regarding rival entry was not taken into account explicitly.

#### REFERENCES

- G. C. Archibald, "The Qualitative Content of Maximizing Models," *J. Polit. Econ.*, Feb. 1965, 73, 27-36.
- D. P. Baron, "On Models of Limit Pricing and Potential Entry," working paper, Northwestern Univ. 1971.
- Y. Barzel, "Excess Capacity in Monopolistic Competition," *J. Polit. Econ.*, Sept./Oct. 1970, 78, 1142-49.
- W. J. Baumol, *Business Behavior, Value and Growth*, New York 1967, ch. 6.
- , "Monopolistic Competition and Welfare Economics," *Amer. Econ. Rev. Proc.*, May 1964, 54, 44-52.
- J. N. Bhagwati, "Oligopoly Theory, Entry Prevention, and Growth," *Oxford Econ. Pap.*, Nov. 1970, 22, 297-310.
- E. H. Chamberlin, *The Theory of Monopolistic Competition*, Cambridge 1962, ch. 5.
- H. Demsetz, "The Nature of Equilibrium in Monopolistic Competition," *J. Polit. Econ.*, Feb. 1959, 67, 21-30.
- F. H. Hahn, "Excess Capacity and Imperfect Competition," *Oxford Econ. Pap.*, Oct. 1955, 7, 229-40.
- R. F. Harrod, "Theory of Imperfect Competition Revised," *Economic Essays*, New York 1952, 139-57.
- J. R. Hicks, "The Process of Imperfect Competition," *Oxford Econ. Pap.*, Feb. 1954, 6, 41-54.
- N. Kaldor, "Market Imperfection and Excess Capacity," *Economica*, Feb. 1935; reprinted in N. Kaldor, *Essays on Value and Distribution*, London 1960, 62-95.
- M. I. Kamien and N. L. Schwartz, "Limit Pricing and Uncertain Entry," *Econometrica*, May 1971, 39, 441-54.
- G. Pyatt, "Profit Maximization and the Threat of New Entry," *Econ. J.*, June 1971, 81, 242-55.
- P. A. Samuelson, "The Pure Theory of Public Expenditure," *Rev. Econ. Statist.*, Nov. 1954, 36, 387-89.

# The Allocation of Transitory Income Among Consumers' Assets

By MICHAEL R. DARBY\*

This article uses an all-capital model to consider the allocation of transitory income among consumers' assets. All transitory income is assumed to be saved. Two basic hypotheses are stated: 1) A fraction of transitory income flows into transitory cash holdings, and these holdings are converted into durable goods and financial assets at a constant rate. 2) The fraction of transitory income immediately used to accumulate durable goods depends on the households' present holdings of transitory assets. The greater the absolute value of the ratio of such assets to total permanent assets, the smaller will be the fraction of transitory income converted into durable goods, and therefore the larger will be the conventionally defined savings ratio.

An error in the standard method of estimating permanent income is reported in the Data Appendix, and a general class of easily reproducible estimators is presented. The increase in money demand resulting from transitory income is put at .4 or .5 of transitory income, and the estimated rate of adjustment of transitory money holdings was about .2 or a bit higher. Hypothesis 2) is strongly supported by the tests conducted. The maximum fraction of nonmonetary transitory savings going into durable goods is put at around .4.

The first hypothesis offers one reason

why the lags in the effect of monetary policy might appear long and variable. The second hypothesis implies that the marginal propensity to spend out of transitory income is variable and near its minimum at the cyclical peak and trough. This weakens the basis of the multiplier analysis, particularly for countercyclical policy recommendations. The model as a whole provides a transmission mechanism for changes in the quantity of money to directly affect spending.

The importance of distinctions between transitory and permanent stocks and flows receives strong confirmation from this study.

The first section outlines the theoretical framework. The second discusses the estimation of the demand for transitory money holdings and presents the empirical results. The third section presents the remaining empirical results. The final section discusses the implications of these results, with emphasis on macro-economic theory.

## I. The All-Capital Model

### *Definitions*

My basic theoretical model is an all-capital world derived from Milton Friedman (1957).<sup>1</sup> The consumer holds his wealth in the forms of money, financial

\* Assistant professor of economics, Ohio State University and visiting assistant professor, University of California, Los Angeles. This paper is derived primarily from my dissertation and was supported by a Federal Deposit Insurance Corporation Graduate Fellowship in banking, finance, and economics. I am especially indebted to Milton Friedman, Marc Nerlove, J. Richard Zecher, P. A. V. B. Swamy, and Arnold Zellner.

<sup>1</sup> Most previous work in this framework has been on the consumption function, but some work has been done on the effects of transitory income upon specific assets, most notably by Gregory Chow (1955), (1960), (1966), Brian Motley, Meyer Burstein (1957), (1960), and Paul Smith. The present article is the first to consider the total allocation of transitory income under a portfolio constraint.

assets,<sup>2</sup> durable goods, and human capital. Permanent income flows from these assets at a permanent (or long-run expected) rate  $r_p$ .<sup>3</sup> These assets are the integral of the past flows of permanent savings.<sup>4</sup>

The concept of dividing an "appropriately" measured income into permanent and transitory components is by now familiar. Transitory income is not true income but merely a shift of expected income between then and now.<sup>5</sup> The receipt of transitory income is a receipt of assets in the current period offset by a liability of equal net present value consisting of income receipts at a lower level in the past and future.<sup>6</sup> In a perfect capital market, permanent consumption, assets, and savings are based upon long-run expectations; so the assets received (or lost) as transitory income would be accumulated over time. These assets would have a long-run expectation of zero, but would only rarely attain that particular figure.

At this point it is necessary to introduce some notation: The subscripts  $P$ ,  $T$ , and  $N$  denote, respectively, the permanent, transitory, and nonmonetary transitory components of the variable which they modify.

$A_t$  = Assets at time  $t$ <sup>7</sup>

$F_t$  = Income at time  $t$

<sup>2</sup> In my view, liability choices are essentially part of the portfolio decision; so liabilities of ultimate wealth holders are treated as negative financial assets.

<sup>3</sup> "These assets" properly refers only to (permanent) assets which are viewed as making up wealth as opposed to (transitory) assets which merely offset an expected future receipt of opposite sign.

<sup>4</sup> Thus permanent income is viewed as an effect of rational decisions based on the individual's possibilities and utility surface.

<sup>5</sup> This differentiates transitory income from windfalls, which are fortuitous, instantaneous changes in wealth, and which are discussed later in this section.

<sup>6</sup> The reverse would occur for a receipt of negative transitory income, of course. The reader can make this qualification, where necessary, in the remainder of this paper.

<sup>7</sup> All stocks and flows are in real terms. All flows in this section are at instantaneous annual rates.

$S_t$  = Savings at time  $t$

$M_t$  = Money holdings at time  $t$

$D_t$  = Durable goods stock at time  $t$

Familiar demand curves exist which determine the allocation to specific assets of the fixed, at any instant of time, total permanent assets. The variables involved are yields on the various assets, permanent income (or assets), expected price changes, and  $r_p$ . The inclusion of age and family status captures the insight of the "life cycle" hypotheses while still allowing for long-run saving.<sup>8</sup> In future sections more explicit assumptions will be necessary.

#### *Hypotheses on the Allocation of $Y_{Tt}$*

Transforming one of the usual identities, we have

$$(1) \quad S_{Tt} = Y_{Tt} - C_{Tt}$$

Transitory consumption,  $C_{Tt}$ , is taken as an independent random disturbance with zero mean. Transitory savings,  $S_{Tt}$ , is therefore the amount to be allocated.

Money is taken to be currency plus demand deposits. For this money, like a consumers' durable good,<sup>9</sup> the return is mainly in the form of a flow of services. Unlike durable goods, the value of money holdings cannot be changed without changing the service flow.<sup>10</sup> The service flow at the permanent level is optimal by utility maximization, but this does not imply that money will therefore be fixed at the permanent level. The flow of services is based on its use in making transactions or, more generally, as a store of value, and these functions require that the level of money fluctuate around the long-run level in order to yield utility. Thus we are in a dynamic framework in which the

<sup>8</sup> In this regard it is important to recall the inclusion of human capital in the individual's portfolio.

<sup>9</sup> On this point, see Chow (1966), pp. 113, 126.

<sup>10</sup> For example, by changing age, durability, or required maintenance.

current desired level depends on past, present, and expected future conditions.

Transitory savings are in the nature of a shock. Taking account of costs of adjustment, it is hypothesized that a large proportion of transitory savings, positive or negative, will first flow into money holdings and then, only gradually, will money holdings be adjusted to the permanent level as the surplus or deficit is eliminated by transfers to or from financial assets or durable goods or transitory savings of opposite sign. This "shock absorber" hypothesis can be written as

$$(2) \quad \frac{dM_{Tt}}{dt} = \beta_1 S_{Tt} + \beta_2 M_{Tt}$$

where  $\beta_1$  is the fraction of transitory savings going directly into cash, and  $\beta_2$  ( $\beta_2 < 0$ ) is the rate at which transitory money holdings are drawn down.<sup>11</sup> Thus, if a level of  $S_{Tt}$ , say  $S_{Tt}^0$ , were maintained for long enough,  $dM_{Tt}/dt$  would approach zero as  $M_{Tt}$  approached  $-(\beta_1/\beta_2)S_{Tt}^0$ .

Define nonmonetary transitory savings as

$$(3) \quad S_{Nt} = S_{Tt} - \frac{dM_{Tt}}{dt}$$

The remaining problem is to determine the allocation of this amount among durable goods, financial assets, and human capital.

Consumers' durable goods essentially yield income to the consumer as a flow of services. Assuming the market works, durable goods stocks of different values can yield close to the same total cost by variations in the mix of interest, depreciation, maintenance, and other costs. Durable goods have high transaction costs for large changes, but gradual changes in

value can be easily accomplished by altering timing of replacement purchases.

Financial assets yield income primarily in the form of pecuniary returns. I assume that at least for some of these assets a yield close to the permanent interest rate can be obtained with small transactions costs.

Human capital investments yield both service flows directly and rental income in the labor market. Human capital is very illiquid as sale is prohibited and depreciation is generally slow.

Human capital seems a poor candidate for transitory investments<sup>12</sup> for the reasons just outlined; so I assume essentially no transitory investment in human capital occurs. Therefore transitory investment is to be apportioned between durable goods and financial assets. Let the fraction of a marginal transitory investment going into durable goods (1—the fraction going into financial assets) be called  $\delta$ , and let  $\delta_0$  be its value if nonmonetary transitory assets,  $A_{Nt}$ , are zero. I hypothesize that  $\delta$  will monotonically decrease from  $\delta_0$  as the absolute value of  $A_{Nt}$  increases. This follows directly from the assumed increasing costs as the level of durable goods diverges further from the optimal long-run (permanent) levels due to increasing costs of substitution in consumption and production<sup>13</sup> as well as increasing liquidity problems, and from the observation that financial assets suffer these losses to a much lesser extent, due to the greater importance of their pecuniary yield and their more liquid nature. Low or high levels of  $|A_{Nt}|$  must be relative to something. In this case  $A_{Pt}$  is the obvious choice, representing the size of the base to which these changes are being made.

<sup>11</sup> These  $\beta$ 's may vary in response to yet other variables (such as in a hyperinflation), but will be taken as constants for the current problem. Note that contrary to Chow (1966) and Motley, the assumption is made—in order to obtain sensible steady-state results—that changes in desired levels are fully adjusted to, but that only a fraction of the transitory portion is adjusted.

<sup>12</sup> "Investment" is used in a strictly microeconomic sense here and should not be confused with the concept of the same name used in the standard macroeconomic model.

<sup>13</sup> Under the usual curvature assumptions on the utility and production functions.

Thus all  $S_{Nt}$  is assumed to be divided between durable goods and financial assets. The hypothesis can be stated as

$$(4) \quad \frac{dD_{Tt}}{dt} = f(A_{Nt}/A_{Pt}) \cdot S_{Nt}$$

$$(5) \quad \frac{df}{d[A_{Nt}/A_{Pt}]} < 0$$

#### *Qualifications to the Model*

Part of what we call transitory income consists really of windfalls, and this should be separated out. If people aren't sure at the time of receipt whether or not wealth has increased, a prediction scheme in which a fraction of transitory assets flows into permanent assets is suitable, and this is basically how we in fact estimate permanent income.

Besides the usual aggregation problems, much transitory income is cancelled out in the aggregate data; so that the focus must be on cyclical problems. This could cause difficulties if the distribution of transitory income across families changes markedly within the sample period. On aggregation, however, we have much less need to worry about nonzero values of  $C_{Tt}$  and can substitute  $Y_{Tt}$  for  $S_{Tt}$ .

The model assumes that any asset allocation within the wealth constraint can be accomplished. This is true for the individual, and I must assume that inventories and production are such that it will also hold for the economy as a whole.

The usual care must be taken in applying the constructs in this model to similarly named data or constructs in income-expenditures models.

## II. Data Series and Money Demand Results

### *Data Series and Estimates*

The period covered in the empirical study is the 80 quarters from 1947-I to 1966-IV. This was the longest period for which all required data series were complete at the start of estimation.

Brief descriptions of the data series used follow.<sup>14</sup> The basic series are:

$P_t$  = Consumer Price Index (1957-1959 base), middle month of quarter.

$P_{Dt}$  = Implicit Personal Consumption Durable Goods Price Index (1958 base), quarterly.

$Y_t$  = Private Income = disposable income + undistributed corporate profits,<sup>15</sup> at seasonally adjusted quarterly rates (SAQR), deflated by  $P_t$ .

$M_t$  = Money supply (currency + demand deposits), seasonally adjusted last month of quarter (SALMQ), deflated by  $P_t$ .

$HM_t$  = High powered money, SALMQ, deflated by  $P_t$ .

$DX_t$  = Personal consumption durable goods expenditures, SAQR.

$RL_t$  = U.S. government long-term bond yield, quarterly average (QA).

$RS_t$  = U.S. Treasury three-month bills, market yield, QA.

The variables  $Y_t$ ,  $M_t$ , and  $HM_t$  are in real terms (deflated by  $P_t$ ) with dimensions of 1958 dollars (per quarter for  $Y_t$ ). The yield variables are in percentage points which have the dimensions of 1/time. The price indices have the dimension t dollars/1958 dollars with 1958 = 1.000.

The stock of durable goods, end of quarter,  $D_t$ , is estimated from the quar-

<sup>14</sup> Sources for the  $P_t$ ,  $P_{Dt}$ ,  $Y_t$ , and  $DX_t$  are the 1966 and 1967 supplements to the *Survey of Current Business*. The  $M_t$  series is from various issues of the *Federal Reserve Bulletin*. The  $HM_t$  series is the source base published in the August 1968 *Review* of the Federal Reserve Bank of St. Louis. Both  $RL_t$  and  $RS_t$  data were taken from the Board of Governors of the Federal Reserve System, *Supplement to Banking and Monetary Statistics, Section 12, Money Rates and Securities Markets* and various issues of the *Federal Reserve Bulletin*. Complete detail is found in the appendix to my dissertation.

<sup>15</sup> Undistributed corporate profits estimate the capital gains that should be included in income. No attempt is made to estimate the service flows direct from assets.

terly durable goods expenditures by application of a perpetual inventory method based on double declining balance depreciation.<sup>16</sup>

$$(6) \quad D_t = .95 \cdot D_{t-1} + DX_t/P_{Dt}$$

Raymond Goldsmith's estimates are used for the 1946 benchmark.

The estimated yield on money, quarterly average,  $RM_t$ , is derived by Benjamin Klein's method:

$$(7) \quad RM_t = (1 - HM_t/M_t)RS_t$$

This is based on the hypothesis that banks avoid the restriction on payment of interest on demand deposits completely. This hypothesis will be tested in connection with other texts.

The computation of permanent income,  $Y_{Pt}$ , and transitory income,  $Y_{Tt}$ , from  $Y_t$  is explained in the Data Appendix.

#### *Empirical Results for Money Demand*

The general problems in aggregation have already been discussed. For the money demand model there is an additional problem. The money stock is held not only by individuals but also by businesses, and no good data exist which estimate the stock held by each. In testing the model, one must view the firms either as holding the cash for their owners or—what is more reasonable—as acting in much the same way as ultimate wealth holders in regard to transitory money balances. I make the latter assumption and apply the model to the available data.

The shock absorber hypothesis of money demand is summarized in equation (2). In order to apply discrete time-series to the continuous time model, it is taken that transitory demand at the end of a quarter will be increased by a fraction of transitory savings in the quarter and that a

fraction of transitory money holdings at the beginning of the period will be worked off over the course of the quarter. Unfortunately there is no series on transitory money stock nor any obvious way to estimate one. However since measured money stock is the sum of its transitory and permanent components, it is possible to test the hypothesis when it is combined with a linear permanent demand function:

$$(8) \quad \begin{aligned} M_t - M_{t-1} = & \beta_1 Y_{Tt} + \beta_2 (M_{t-1} - M_{Pt-1}) \\ & + M_{Pt} - M_{Pt-1} + \epsilon_t, \end{aligned}$$

where  $\epsilon_t$  is a random term admitted to the transitory demand to allow for shocks and discrepancies and where  $Y_{Tt}$  has been substituted for  $S_{Tt}$  as previously discussed. Simplifying,

$$(9) \quad \begin{aligned} M_t = & \beta_1 Y_{Tt} + (1 + \beta_2)M_{t-1} + M_{Pt} \\ & - (1 + \beta_2)M_{Pt-1} \end{aligned}$$

I assume that permanent money demand is based on the linear demand function:

$$(10) \quad \begin{aligned} M_{Pt} = & \beta_3 + \beta_4 \cdot Y_{Pt} + \beta_5 \cdot RL_t \\ & + \beta_6 \cdot RS_t + \beta_7 \cdot RM_t \end{aligned}$$

Substituting equation (10) into equation (9) and simplifying,

$$(11) \quad \begin{aligned} M_t = & \beta_3(1 - \beta_8) + \beta_1 Y_{Tt} + \beta_8 M_{t-1} \\ & + \beta_4 Y_{Pt}^* + \beta_5 RL_t^* + \beta_6 RS_t^* \\ & + \beta_7 RM_t^* + \epsilon_t, \end{aligned}$$

where  $\beta_8 = 1 + \beta_2$  and the asterisked variables are computed by substituting for  $X$  in

$$(12) \quad X_t^* = X_t - \beta_8 X_{t-1}$$

By the usual Taylor series justification, this non-linear equation may be estimated by ordinary least squares iterated on the estimated values of  $\beta_8$ .<sup>17</sup> If the equation

<sup>16</sup> This assumes a ten-year average life of durable goods. See Laurits Christensen and Dale Jorgenson, pp. 294-97.

<sup>17</sup> The problem is essentially in the multiplicative constraints on the coefficients in equation (9). The differencing of the permanent variables as in equations (11)

TABLE 1—REGRESSION COEFFICIENTS: MONEY DEMAND EQUATION DEPENDENT VARIABLE:  $M_t$ 

		Constant	$Y_{Tt}$	$M_{t-1}$	$Y_{Pt}^*$	$RL_t^*$	$RS_t^*$	$RM_t^*$	$R^2$	$SEE$
1	1947-II–	24.901	.3797	.8047	.1938	–1.1075	–36.246	56.745	.94701	1.18927
	1966-IV	(5.6001)	(.09602)	(.04327)	(.0757)	(1.5583)	(9.1490)	(14.413)		
		4.4466	3.9542	18.596	2.5608	–.7107	–3.9618	3.9371		
2	1953-I–	30.118	.5651	.7674	.1599	–.3628	–31.935	49.104	.96619	.84158
	1966-IV	(9.3135)	(.1089)	(.07158)	(.06813)	(1.2513)	(8.7407)	(13.647)		
		3.2338	5.1881	10.721	2.3466	–.2900	–3.6536	3.5981		
3	1947-II–	49.36	.5664	.6040	.2483	–1.505	–21.58	33.41	.86297	2.02406
	1966-IV	(8.714)	(.1118)	(.0690)	(.0064)	(1.189)	(7.992)	(12.63)		
	GLS	5.664	5.068	8.758	3.741	–1.266	–2.700	2.645		

Note: The standard errors are given in parentheses below the coefficients and the  $t$ -values are given below the standard errors.

iterates, it has the usual properties of a least squares regression. Note however that there are still problems of the lagged dependent variable which implies that the estimates have only large-sample justification and possible problems from autocorrelation of the residuals.

Certain predictions can be made about the size and sign of the values of the regression coefficients. The coefficient  $\beta_1$  estimates the fraction of transitory income flowing into cash balances and not reinvested during the quarter; so  $\beta_1$  should be in the range 0 to 1, with the middle more likely than either extreme. The coefficient  $\beta_2$  is restricted to the range from 0 to  $-1$ , but should be no slower than people adjust their expectations ( $-.1$ ); also  $\beta_2 < -.5$  is unlikely since it does not reflect any outflow in the first quarter at a possibly higher rate. Therefore the estimated range of  $\beta_3 = 1 + \beta_2$  is from .5 to

.9. The coefficient  $\beta_4$  should be positive and, under the simplest form of the quantity theory, should equal about four times the fraction of a year's income held as money,<sup>18</sup> around 1.7. The coefficients  $\beta_5$  and  $\beta_6$  should be negative, as these interest rates reflect yields on alternatives to money holdings. If the hypothesis that banks avoid the prohibition of payment of interest on demand deposits is true,  $\beta_7$  should be positive and greater in absolute value than either  $\beta_5$  or  $\beta_6$ , which are for cross-yields.

Table 1 reports the results of running the regressions for equation (11). Besides the regression for the whole period, there is also an equation for 1953-I through 1966-IV and a generalized least squares (GLS) version of regression.

Number 1 is the key regression. The three interest rate coefficients are entirely consistent with Klein's hypothesis.<sup>19</sup> The

and (12) allows the use of a simple iterative method of solution: 1) Choose an initial value,  $\beta_8^0$  of  $\beta_8$  (I estimated an unconstrained version of equation (9) for this but the convergence was rapid so that 1.0 would have been as good); 2) compute the asterisked variables in equation (11) using  $\beta_8^0$  and equation (12); 3) run the OLS regression for equation (11) using these variables; and 4) replace the value  $\beta_8^0$  with the regression coefficient of  $M_{t-1}$  in step 3) and repeat steps 2) through 4) until the absolute value of the change in the estimated  $\beta_8$  is less than some small tolerance (here .0001).

<sup>18</sup> Recall that flows are expressed as quarterly rates.

<sup>19</sup> Converting these coefficients to elasticities of demand for money at the sample mean, the figures are  $-.01$  with respect to  $RL_t$ ,  $-.61$  for  $RS_t$ , and  $+.60$  for  $RM_t$ ; so proportionate increases in all three interest rates would leave money demand virtually unaffected. I also ran the basic equations deleting the  $RM_t^*$  term; there were no significant changes in the noninterest rate coefficients, but the  $RS_t$  coefficient went to an insignificant  $-.31$  with the  $RL_t$  coefficient virtually unchanged. The elasticities are  $-.01$  for  $RL_t$  and  $-.005$  for  $RS_t$ . It would appear that earlier studies which found a sig-

estimate of  $\beta_1$  is in the expected range and significantly different from zero. The estimate of  $\beta_8$  is in the high end of the predicted range, indicating a fairly slow adjustment of transitory money balances.<sup>20</sup> The estimate of  $\beta_4$  is much lower than predicted by the simple quantity theory.

The lower than mid-range estimates of  $\beta_1$  and  $\beta_8$  might be attributed to the increased moneyiness of securities pegged by the Federal Reserve before 1953 relative to the later part of the period.<sup>21</sup> Regression 2 was run to test this possible change in the underlying model. The estimates of  $\beta_1$  and  $\beta_8$  move in the expected direction, but the null hypothesis that the 1947-II<sup>22</sup> to 1952-IV sample was drawn from the same population as the 1953-I through 1966-IV sample cannot be rejected at the .95 level of significance.<sup>23</sup> Therefore the 1953-I through 1966-IV regressions cannot be regarded as other than suggestive.

The lower coefficient estimated for permanent income can be attributed to the linearity restriction on the permanent money demand function, to the hypothesis that unitary income elasticity is far from correct, to the hypothesis that a

broader definition of money is required, or to the particular period studied.<sup>24</sup> The variable  $Y_{Pt}^*$  has very little variation from a trend, so it would be difficult to obtain estimates of the separate effects of changes in permanent income. Hence it is not clear precisely what information is contained in the estimate of  $\beta_4$ .

The shock absorber model is strongly supported by the data. Transitory income increases the demand for money by about 40 percent of the amount of transitory income, and transitory money balances are worked off at a rate of about 20 percent per quarter. There is suggestive evidence that both these percentages have increased since the 1940's. All the estimates were significantly different from both ends of the range permitted within the model.

### III. Durable Goods Demand Results

#### *Estimation Procedure*

It is desired to construct a test of the hypothesis of equations (4) and (5) on the allocations of  $S_{Nt}$  between durable goods and financial assets. In this case also there are data problems. Since there is no  $D_{Tt}$  series, I must again assume a permanent demand function in linear form:

$$(13) \quad D_{Pt} = \gamma_1 + \gamma_2(P_{Dt}/P_t) + \gamma_3 Y_{Pt} + \gamma_4 RL_t$$

This is essentially Chow's (1960) hypothesis, with the interest rate term added because durables are viewed as competing with other assets for a share of the whole.<sup>25</sup>

<sup>24</sup> I am indebted to Friedman for the point that in this period money demand was adjusting slowly to its long-run demand from the high levels caused by the uncertainties and illegal activities of the war prior to, and during, the early part of the period, while alternatives to money, such as time deposits and savings and loan shares, were growing rapidly. Rerunning the regressions for  $M_{2t} = M_t + \text{time deposits}$  and for 1961-I through 1966-IV increases the estimate of this coefficient, but not enough to solve all the problem.

<sup>25</sup> Although the inclusion of  $RS_t$  might also seem to be an improvement, the life of durable goods appears to be too long for this rate to have any effect; the inclusion of  $RS_t$  in the regression has essentially no effect.

nificant *net* effect of interest rate changes suffered from specification bias, with the interest rate variable serving as a proxy for the omitted variable transitory income.

<sup>20</sup> It is possible that this high value of  $\beta_8$  may in part be due to autocorrelation in the residuals. E. Malinvaud (p. 472) summarized the evidence and suggests ordinary least squares in this type of model (even without the complication of the non-linear constraints, because of the relatively large errors inherent in estimating the correlation coefficient  $\rho$ ). The main effect of autocorrelation is to bias the estimated standard errors downward; so somewhat higher standards of significance are advisable in considering the regression coefficients presented here. Nevertheless, I did run a generalized least squares version of regression 1 (using an estimated  $\rho$  of .5341). The estimate of  $\beta_8$  decreased to a bit over .6 and of  $\beta_1$  increased into the .5 range, but there were no other interesting changes. Little reliance should be put on these estimates, but they are presented as regression number 3.

<sup>21</sup> See Friedman and Anna Schwartz, p. 625.

<sup>22</sup> The first quarter was lost due to lagged variables.

<sup>23</sup> The  $F(7, 65)$  statistic of 1.84 is such that the hypothesis would just be rejected at the .90 level of significance.

I chose to represent the variation in the fraction of  $S_{Nt}$  going into  $D_{Tt}$  by the specific form:

$$(14) \quad \Delta D_{Tt} = \gamma_5 e^{-(\eta A_{Nt}/A_{Pt})^m} \cdot S_{Nt} + \epsilon_t,$$

where  $m=4$ ,<sup>26</sup>  $\epsilon_t$  is a random error, and  $\gamma_5$  and  $\eta$  are parameters to be estimated. The coefficient of  $S_{Nt}$  has the desired properties of having a maximum value ( $\gamma_5$ ) when  $A_{Nt}=0$  and decreasing with increasing  $|A_{Nt}/A_{Pt}|$ . Estimation of the  $\eta$  parameter allows for search over a wide range of rapidity of that decrease.

Unfortunately, data series are lacking for the three variables on the right-hand side of equation (14). The replacement of  $A_{Pt}$  by  $Y_{Pt}$  is straightforward.<sup>27</sup> I use  $Y_{Tt}$  as a proxy for  $S_{Nt}$ ,<sup>28</sup> and  $A_{Tt}$  for  $A_{Nt}$ . The estimate of  $A_{Tt}$  is based on the prediction scheme discussed in the last part of Section I:

$$(15) \quad A_{Tt} = (1 - b)A_{Tt-1} + Y_{Tt}$$

The choice of the parameter  $b$  is the same as for permanent income, or .1. However an initial value,  $A_{T,0}$ , is required. Even the sign of  $A_{T,0}$  is an open question. Some would argue that people were not yet recovered from paying for the war at the end of 1946; while others would suggest that excess assets were accumulated during the war. Lacking a direct estimate,  $A_{T,0}$  will be left as a parameter to be estimated.

Now substituting in equation (14) and

<sup>26</sup> A check of other values of  $m$  showed  $m=2$  to display too little variation; the results were substantially the same for  $m=6$  as for  $m=4$ .

<sup>27</sup> Since  $Y_{Pt} = r_P \cdot A_{Pt}$  for annual rates, this will merely decrease the value of  $\eta$  by a factor of  $r_P/4$ .

<sup>28</sup> Changes in  $M_{Tt}$  including business balances would not be really appropriate for estimating  $S_{Nt} = Y_{Tt} - \Delta M_{Tt}$ . Nevertheless, I also tried estimates of  $S_{Nt}$  as  $Y_{Tt}$  less an estimate of  $\Delta M_{Tt}$  (based on the results of the previous section) and on changes in my estimate of  $A_{Tt}$ . However since regression 1 was primarily carried out to test the money demand hypothesis, the  $\Delta M_{Tt}$  estimates introduced the autocorrelation difficulties discussed in fn. 20, while the  $\Delta A_{Tt}$  series was dominated by trend components; so neither attempt led to good results.

adding the result to the first difference form of equation (13), the equation to be estimated is:

$$(16) \quad \Delta D_t = \gamma_1^* + \gamma_2 \cdot \Delta(P_{Dt}/P_t) \\ + \gamma_3 \cdot \Delta Y_{Pt} + \gamma_4 \cdot \Delta RL_t \\ + \gamma_5 \cdot e^{-(\eta \cdot A_{Tt}/Y_{Pt})^4} \cdot Y_{Tt} + \epsilon_t,$$

where the constant  $\gamma_1^*$  should be 0, but is included to remove any linear trend. This equation can be estimated by non-linear least squares over an  $\eta$  and  $A_{T,0}$  search grid.<sup>29</sup>

### *Estimated Durable Goods Demand Equations*

A nonzero  $\gamma_1^*$  would indicate the presence of a linear trend. By simple price theory,  $\gamma_2$  and  $\gamma_4$  should be negative, while  $\gamma_3$  should be positive. The fraction  $\gamma_5$  is restricted to the range from 0 to 1. If it is not significantly different from the marginal propensity to consume, say .9, and  $\eta=0$  is accepted, then a simple func-

<sup>29</sup> An approximate test of whether the inclusion of the exponential factor in the  $\gamma_5$  term is significant can be based on the likelihood ratio. The null hypothesis is that the fraction of transitory income flowing into durable goods is constant regardless of  $A_{Tt}/Y_{Pt}$ , or simply  $\eta=0$ . The alternative hypothesis is that the fraction is determined as in equation (14), or  $\eta \neq 0$ . Assuming independent normally distributed disturbances, the likelihood ratio is

$$\lambda = (SS/SS_0)^{T/2}$$

where  $SS$  is the sum of the squared residuals of the unconstrained estimate,  $SS_0$  is the sum of squares of the estimate made assuming the null hypothesis to be true, and  $T$  is the number of observations. (Here 79; one observation is deleted due to taking first differences.) This is known to be distributed, if the null hypothesis is true and as  $T \rightarrow \infty$  according to  $-2 \cdot \log \lambda \sim \chi^2(k)$ , where  $k$  is the number of restrictions, in this case 1 (see Maurice Kendall and A. Stuart, pp. 230-31). Therefore the test of the significance of this alternative hypothesis is based on comparison of the statistic  $-T \cdot \log(SS/SS_0)$  with the  $\chi^2(1)$  distribution. The specification (14) is only one of many possible specifications so that acceptance of the null hypothesis would not be conclusive evidence against the hypothesis of equations (4) and (5), except as this specification is a good approximation for most other specifications of this hypothesis, but rejection of the null hypothesis would offer strong evidence in support of the hypothesis of those equations.

TABLE 2—REGRESSION COEFFICIENTS: DURABLE GOODS DEMAND EQUATION, DEPENDENT VARIABLE:  $\Delta D_t$ 

	Constant	$\Delta\left(\frac{P_{Dt}}{P_t}\right)$	$\Delta Y_{Pt}$	$\Delta RL_t$	$\gamma_5$	$\eta$	$A_{T,0}$	$R^2$	SEE
4	1.5041 (.1674) 8.9863	-18.997 (6.5799) -2.8871	1.1807 (.2133) 5.5355	-1.1489 (.5538) -2.1522	.4014 (.05182) 7.7455	6.4	-20	.778	.527
5	1.5144 (.1983) 7.6385	-22.906 (7.3573) -3.1134	1.1872 (.2534) 4.6841	-1.0710 (.5935) -1.8046	.3077 (.05181) 5.9400	[0]	—	.728	.584

Note: The standard errors are given in parentheses below the coefficients and the  $t$ -values are given below the standard errors.

tion with durable goods expenditures as consumption is satisfactory. The search grid was a matrix in which  $A_{T,0}$  varies from -\$50 billion to \$50 billion in steps of \$10 billion, and  $\eta$  varies from 0 to 100,<sup>30</sup> ultimately in steps of .1.

The non-linear least squares regression for equation (16) is presented in Table 2 as regression number 4. There is a significant trend,<sup>31</sup> but otherwise the coefficients are as expected. The estimates of  $\eta$  and  $A_{T,0}$  are 6.4 and -\$20 billion, respectively.

Regression 5 was run under the assumption  $\eta=0$  to test the significance of the exponential part of the  $\gamma_5$  term,<sup>32</sup> and the null hypothesis was easily rejected in favor of the alternative of  $\eta \neq 0$ .<sup>33</sup> This offers strong support to the hypothesis that a decreasing fraction of transitory income goes into durable good as increasing absolute values of  $A_{Nt}/A_{Pt}$ .

These results strongly confirm the im-

portance of taking separate account of durable goods expenditures, as the marginal propensity to spend out of transitory income on durables is at most about .5, and at times less. Paul Smith reported results of a two-stage least squares estimation that indicated that the marginal propensity to spend out of transitory income on nondurable goods and services is about zero. I checked directly on three alternative simple consumption functions, with  $Y_{Pt}$  and  $Y_{Tt}$  independent variables, and as dependent variables, alternatively:

$C1_t = \text{Personal Consumption Expenditures, } SAQR$

$C2_t = C1_t - \text{Personal Consumption of Durable Goods, } SAQR$

$C3_t = C2_t - \text{Personal Consumption of Clothing and Shoes, } SAQR^{34}$

The regression results are presented in Table 3. Note that regression 6 is misspecified under the hypothesis of this paper, but should provide a reasonable estimate of the average marginal propensity to spend out of transitory income. The other regressions indicate, not surprisingly, that inventory changes are important for clothes and shoes, as for durable goods as

<sup>30</sup>  $\eta$  is basically a scaling factor and sign makes no difference in the value of the expression; so only positive values were considered. A maximum value of 100 appeared to be much more than sufficient.

<sup>31</sup> The significant constant is probably due to trends in tastes, an income elasticity of demand for durable goods greater than unity, or too slow a rate of depreciation, or some combination of these factors.

<sup>32</sup> This means an OLS regression on equation (16) with the penultimate term changed to  $\gamma_5 Y_{Tt}$ .

<sup>33</sup> The  $\chi^2$  test statistic (see fn. 29) is equal to 16.1. The .95 point of the  $\chi^2(1)$  distribution is 3.84 and the .999 point is 10.8.

<sup>34</sup> All consumption expenditures are in terms of 1958 dollars. The series on seasonally adjusted real expenditures on clothing and shoes was generously provided by the Office of Business Economics, U.S. Department of Commerce.

TABLE 3—SIMPLE CONSUMPTION FUNCTIONS

	Dependent Variable	Constant	$Y_{Pt}$	$Y_{Tt}$	$R^2$	SEE
6	$C_{1t}$	-1.640 (.3807) -4.308	.9026 (.00456) 198.0	.5005 (.04028) 12.43	.99812	.66649
7	$C_{2t}$	1.867 (.2677) 6.975	.7318 (.00320) 228.3	.1226 (.02832) 4.328	.99856	.46854
8	$C_{3t}$	.4280 (.3008) 1.423	.6716 (.00360) 186.5	.04306 (.03183) 1.353	.99783	.52659

Note: The standard errors are given in parentheses below the coefficients and the  $t$ -values are given below the standard errors.

defined by the Office of Business Economics, and account for most of the remaining effect of  $Y_{Tt}$ . The final regression is in almost complete conformity with Friedman's hypothesis, even to the insignificant constant term.

#### IV. Implications and Conclusions

The results strongly support Friedman's permanent income approach. Both the direct hypothesis of a zero effect for transitory income on consumption, and derived hypotheses on the demand for money and durable goods received confirmation from the analysis of the aggregate data for the postwar period.

The marginal propensity to spend out of private income derived from the results is of particular interest. Permanent income is consumed according to a proportionality factor; so we can take, for purposes of cyclical analysis, this marginal propensity to be this factor of proportionality multiplied by the increase in permanent income due to current income<sup>35</sup> plus the marginal propensity to spend transitory income.<sup>36</sup> It was shown that transitory income is spent, in the usual

sense, only to build up the inventories of durable or semi-durable goods. It was also shown that the fraction of transitory income so spent varies with the absolute value of the ratio of transitory assets to permanent assets. Now this absolute value moves cyclically, with peaks in its value near the peaks and troughs of the business cycle. Thus the marginal propensity to spend also fluctuates and is near minimum at these points.<sup>37</sup> Therefore the standard macro-economic models would be at their weakest near these points, due to the apparent increases in the savings ratio defined in terms of consumption goods expenditures.

It is possible to give more precise numerical values to the quantities involved. Taking the total inventory effect to follow the pattern of the dominant durable goods portion, the maximum value, for zero transitory assets, of the marginal propensity to spend current income would be

<sup>37</sup> There are eight National Bureau of Economic Research (NBER) reference cycle peaks and troughs in the period covered. The average reduction in the marginal propensity to spend on durable goods out of transitory income was to 72 percent of the maximum propensity, but the range was from 8 to 100 percent of the maximum. For 1966-IV the reduction was to 65 percent, although this was not labeled a peak by the NBER. Thus, the variation has economic, as well as statistical, significance.

<sup>35</sup> That is, about  $.9 \times .1 = .09$ .

<sup>36</sup> Although changes in consumers' inventories are not consumption, they do provide, in the short run, changes in the aggregate demand for goods and services.

about .74.<sup>38</sup> The reductions due to higher levels of the absolute value of transitory assets are large enough to reduce the average, over time, marginal propensity to spend current income to about .59. At a cyclical peak or trough this figure would typically be about .56,<sup>39</sup> with further decreases in the quarters just past the peak or trough.<sup>40</sup>

The money demand hypothesis provides an explanation for the lag in the effect of monetary policy. Consider a change in monetary policy which leaves the community, through a portfolio shock, with larger cash balances than planned. These extra cash balances are transitory money balances, which will be drawn down only gradually, thus delaying the full effect of the policy change. At the quarterly rate of .2 it would take about three quarters for half of the effect of the policy change to be felt, which is close to Friedman's (1961) results on the lag. There are however two (offsetting) further factors to be considered. Suggestive evidence indicated that the adjustment coefficient is higher than .2, at least in the post-Accord period, and may be in the range from .25 to .4 which would shorten the period for half the effect to be felt to two or even one-and-a-half quarters. On the other hand, the short-run increase in income from the working down of money balances is an increase in transitory income and hence absorbs part of the money balances, thus increasing the lag to some extent.

The previous case was for zero initial transitory money balances. If the Federal Reserve was engaged in a change of money

tary policy, initial transitory money balances would be of opposite sign and the magnitude of the effect would depend on the relative size of the policy change to the transitory money balances built up under the preceding policy. This factor may explain the observed variability in the initial effect of a policy change of a particular size, since the initial effect will be generally weaker the longer the preceding opposite policy was maintained.

Another related result of interest to monetary theorists is the fact that interest rate changes had generally insignificant net effects, earlier estimates of significant effects apparently being due to failure to take separate account of permanent and transitory income effects.

Macro-economic modelers should have no difficulty in adapting these results to their particular model. Usual care should be taken in going from estimates of stocks to expenditures, and the concept of private income is new to some models, but generally there are no difficulties. Different estimation techniques might be more appropriate in a modeling, as opposed to testing, context, however.

The most general conclusion from the results of this study is that permanent income and transitory income have separate effects on the demands for assets, the precise nature of the effects depending on the nature of the asset. In particular, practically all previous studies have been concerned with the question: "Does permanent or measured income explain best the demand for this asset?" It is now clear that this question is entirely inappropriate. In general, both have effects.<sup>41</sup> An example is the controversy over whether, abstracting from other variables, money demand is a constant fraction of permanent or measured income. Actually it

<sup>38</sup> The ratio of the  $\gamma_6$  estimates in regressions 4 and 5 (see Table 2) is 1.3, and the coefficient of  $Y_{T1}$  in regression 6 (see Table 3) is .5; so the inventory change response would be put at  $1.3 \times .5 = .65$ . Adding the .09 permanent income effect yields .74.

<sup>39</sup> Computed as  $.65 \times .72 + .09 = .56$ .

<sup>40</sup> Since transitory income is still positive (negative) and adding to the absolute value of transitory assets in the period just following a peak (trough).

<sup>41</sup> Although the appropriate concepts are permanent income and transitory income (measured less permanent income).

seems to be a constant times permanent income plus a (different) constant times transitory income.

### V. Summary and Concluding Remarks

This article investigated the allocation of the transitory component of income among the various classes of consumers' assets. Particular emphasis was placed on the effects of cyclical variations. The analysis and empirical results led to new and interesting conclusions on the marginal propensity to spend and the question of "long and variable lags" in the effect of monetary policy.

Areas for future study would be to apply the model to panel data, extend the period studied, consider whether housing might better be treated as a consumers' durable good, estimate cash holdings of the personal sector, and study the formation of and adjustments in the expectations of the long-run interest rate.

### DATA APPENDIX

#### *Estimation of Permanent Income*

Practically all investigators of permanent income have derived their estimates as

$$(A.1) \quad Y_{Pt} = b \cdot Y_t + (1 - b + c) Y_{Pt-1}$$

where  $b$  is a parameter to be or previously estimated, and  $c$  is the estimate,  $\beta_2$ , of the trend coefficient  $\beta_2$  in the regression

$$(A.2) \quad \log Y_t = \beta_1 + \beta_2 \cdot t + \epsilon_t$$

The initial value of  $Y_{Pt}$  is  $e^{\hat{\beta}_1}$ . This estimate is derived from a continuous time adaptive expectations model. In discrete time, however, what is desired is

$$(A.3) \quad Y_{Pt} = b(Y_t - Y_{Pt}^e) + Y_{Pt}^e,$$

where  $Y_{Pt}^e$  is the expected value in  $t-1$  of permanent income in  $t$ . With a linear trend, this is simply  $(1+c)Y_{Pt-1}$ . So equation (A.3) can be rewritten in this case as

$$(A.4) \quad Y_{Pt} = b \cdot Y_t + (1 - b)(1 + c) Y_{Pt-1},$$

which is the linear adaptive expectations

estimator and differs from the estimator (A.1).<sup>42</sup>

Consider the possibility of a non-linear trend, say due to increasing productivity of capital. That is  $\hat{\beta}_3$  is significant in the regression

$$(A.5) \quad \log Y_t = \beta_1 + \beta_2 \cdot t + \beta_3 \cdot t^2 + \epsilon_t$$

Then differentiation yields

$$(A.6) \quad Y_{Pt}^e = (1 + \hat{\beta}_2 + 2 \cdot \hat{\beta}_3 \cdot t) Y_{Pt-1}$$

The quadratic trended adaptive expectations estimator is

$$(A.7) \quad Y_{Pt} = b \cdot Y_t + (1 - b)(1 + \hat{\beta}_2 + 2 \cdot \hat{\beta}_3 \cdot t) Y_{Pt-1}$$

The extension to any other order trend is straightforward, but for longer periods it might be more appropriate to adjust continuously a linear trend estimated over the previous ten or fifteen years.

The  $\hat{\beta}_3$  coefficients were highly significant for the postwar private income series; so the quadratic trended estimators were used in this paper.<sup>43</sup> Aggregate  $Y_{Pt}$  is computed as

$$(A.8) \quad Y_{Pt} = .1 \cdot Y_t + .9(1.006169 + .00006384 \cdot t) Y_{Pt-1}$$

where in 1947-1  $t$  is 1 and where  $b$  is taken to be .1.<sup>44</sup> The per capita  $Y_{Pt}$  is computed as

<sup>42</sup> The error term  $bc$  is unlikely to change any conclusions from previous studies because of its small magnitude and because permanent income series were usually not derived from computing the trend separately; when the weights  $b$  and  $c$  are estimated simultaneously, the result will be correct though the amount of the trend will be misstated.

<sup>43</sup> The  $t$ -value for the  $t^2$  term was 5.0. Experiments with the linear trended estimator confirmed the superior explanatory power of the quadratic trended estimator, but the differences were not such that any major proposition was dependent upon the choice of estimator.

<sup>44</sup> A quarterly  $b$  of .1 is approximately equal to an annual  $b$  of .35 which is about the middle of the estimates obtained in the various consumption studies. The fact that  $Y_{Pt}$  estimates are used in various phases of this study makes it impossible to choose  $b$  on the basis of best fit in any one equation. A check on robustness over a reasonable range of  $b$ s was made on some of the results; the exact value of  $b$  did not appear to be critical and .1 appeared to be a very good estimate of the actual value.

$$(A.9) \quad Y_{Pt} = .1 \cdot Y_t + .9(1.001521 + .00007624 \cdot t) Y_{Pt-1}$$

The initial values are 58.52 and 411.3, respectively. In a table of data not printed with this paper,<sup>45</sup> I report the values of  $Y_{Pt}$  for aggregate data and also the more customary linear trended estimate  $Y_{Pt}^A$ :

$$(A.10) \quad Y_{Pt}^A = .1 \cdot Y_t + .9078786 \cdot Y_{Pt-1}^A$$

with the initial value of 56.49.

Transitory income series are computed as simply

$$(A.11) \quad Y_{Tt} = Y_t - Y_{Pt}$$

#### *Selected Data Series*

Besides  $Y_{Pt}$  and  $Y_{Pt}^A$ , I have also computed aggregate series of  $Y_t$  (private income),  $D_t$  (durable goods stock),  $F_t$  (financial assets), and  $RM_t$  (yield on money).<sup>46</sup> Where  $YN_t$  is nominal private income,  $PO_t$  is the personal outlays series,  $SAQR$ ,  $\Delta MN_t$  is the change in nominal money supply for the quarter, and  $PI_t$  is the Private Investment Price Index (1958 base), quarterly,  $F_t$  is computed as

$$(A.12) \quad F_t = F_{t-1} + (YN_t - PO_t - \Delta MN_t)/PI_t$$

$F_0$  at the end of 1946 is estimated from Goldsmith as 978.42. Note that all the money stock is subtracted from  $F_t$  and the business portion must be added back for some uses.

#### REFERENCES

- A. Ando and F. Modigliani, "The 'Life Cycle' Hypothesis of Saving: Aggregate Implications and Tests," *Amer. Econ. Rev.*, Mar. 1963, 53, 55-84.
- W. J. Baumol, "The Transactions Demand for Cash: An Inventory Theoretic Approach," *Quart. J. Econ.*, Nov. 1952, 66, 545-56.
- K. Brunner and A. H. Meltzer, "Comment on the Long-Run and Short-Run Demand for Money," *J. Polit. Econ.*, Nov. 1968, 76, 1234-40.
- M. L. Burstein, "The Demand for Household Refrigeration in the United States," unpublished doctoral dissertation, Univ. Chicago 1957.
- , "The Demand for Household Refrigeration in the United States," in A. C. Harberger, ed., *The Demand for Durable Goods*, Chicago 1960, 97-145.
- V. K. Chetty, "On the Long-Run and Short-Run Demand for Money," *J. Polit. Econ.*, Nov. 1969, 77, 921-31.
- G. C. Chow, "The Demand for Automobiles in the United States," unpublished doctoral dissertation, Univ. Chicago 1955.
- , "Long-Run and Short-Run Demand for Money: Reply and Further Note," *J. Polit. Econ.*, Nov./Dec. 1968, 76, 1240-43.
- , "On the Long-Run and Short-Run Demand for Money," *J. Polit. Econ.*, Apr. 1966, 74, 111-31.
- , "Reply: A Note on the Estimation of Long-Run Relationships in Stock Adjustment Models," *J. Polit. Econ.*, Nov. 1969, 77, 932-36.
- , "Statistical Demand Functions for Automobiles and Their Use in Forecasting," in A. C. Harberger, ed., *The Demand for Durable Goods*, Chicago 1960, 147-78.
- L. R. Christensen and D. W. Jorgenson, "The Measurement of U.S. Real Capital Input, 1929-1967," *Rev. Income Wealth*, Dec. 1969, 15, 293-320.
- M. R. Darby, "The Dynamics of the Allocation of Transitory Income Among Consumers' Assets," unpublished doctoral dissertation, Univ. Chicago 1970.
- M. Friedman, "The Demand for Money: Some Theoretical and Empirical Results," *J. Polit. Econ.*, Aug. 1959, 67, 327-51.
- , "The Lag in the Effect of Monetary Policy," *J. Polit. Econ.*, Oct. 1961, 69, 447-66.
- , "Savings and the Balance Sheet," *Bull. Oxford Univ. Inst. Econ. Statist.*, May 1957, 19, 125-36.

<sup>45</sup> The data will be supplied by the author upon request.

<sup>46</sup>  $Y_t$ ,  $Y_{Pt}$ , and  $Y_{Pt}^A$  are in billions of 1958 dollars per quarter (multiply by 4 for annual rates);  $D_t$  and  $F_t$  are in billions of 1958 dollars at the end of quarter;  $RM_t$  is in percentage points and a quarterly average. Complete source detail on these series is located in the appendix to my dissertation.

- , *Studies in the Quantity Theory of Money*, Chicago 1956.
- , *A Theory of the Consumption Function*, Princeton 1957.
- and A. J. Schwartz, *A Monetary History of the United States, 1867–1960*, Nat. Bur. Econ. Res. Stud. in Business Cycles, Vol. 12, Princeton 1963.
- R. W. Goldsmith, *The National Wealth of the United States in the Postwar Period*, Nat. Bur. Econ. Res. Stud. in Capital Formation and Financing, Vol. 10, Princeton, 1962.
- M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics*, Vol. II, New York 1961.
- B. Klein, "The Payment of Interest on Commercial Bank Deposits and the Price of Money: A Study of the Demand for Money," unpublished doctoral dissertation, Univ. Chicago 1970.
- E. Malinvaud, *Statistical Methods of Econometrics*, Chicago 1966.
- F. Modigliani and R. Brumberg, "Utility Analysis and the Consumption Function: An Interpretation of Cross-Section Data," in K. E. Kurihara, ed., *Post Keynesian Economics*, New Brunswick 1954, 388–436.
- B. Motley, "A Demand-for-Money Function for the Household Sector—Some Preliminary Findings," *J. Finance*, Sept. 1967, 22, 405–18.
- P. E. Smith, "The Demand for Durable Goods: Permanent or Transitory Income?," *J. Polit. Econ.*, Oct. 1962, 70, 500–04.
- J. Tobin, "The Interest Elasticity of Transactions Demand for Cash," *Rev. Econ. Statist.*, Aug. 1956, 38, 241–47.
- Fed. Reserve Bank St. Louis Rev., Aug. 1968.
- U.S. Board of Governors of Federal Reserve System, *Fed. Reserve Bull.*, various issues.
- , *Supplement to Banking and Monetary Statistics, Section 12, Money Rates and Securities Markets*.
- U.S. Office of Business Economics, *Surv. Curr. Bus.*, supplements, 1966, 1967.

# COMMUNICATIONS

## Duality and the Many Consumer's Surpluses

By EUGENE SILBERBERG\*

The concepts of duality and indirect utility functions have received renewed interest in recent years owing, in part, to a series of articles by Hendrik Houthakker (1960, 1965) and Paul Samuelson (1965). These discussions stemmed largely from the econometric problem of specifying demand relations. It will be shown that the above concepts provide an interesting and informative way of reformulating the consumer's surplus (or, more aptly, the consumer's surpluses) problem. It is my contention that despite the enormous literature in this area, including extensive analyses by Samuelson (1942, 1947) and John Hicks (1946, 1959), the basic nature of the problem has never been adequately delineated.<sup>1</sup>

The principal reason for the substantial and protracted interest in the concept of a consumer's surplus seems to be the desire to associate changes in monetary (i.e., money income) values with changes in the level of utility, to permit welfare judgments concerning alternative economic equilibria. But a constant source of frustration is that such "welfare" measures are, at best, ambiguous

and normative. This has led to a proliferation of consumer's surpluses, each designed to answer specific questions about the consumer's indifference map. Despite this profusion, one still finds in the literature the single phrase "welfare loss" applied to monopolies, taxes, and so on.<sup>2</sup>

The purpose of this paper is to present a unifying treatment of the many consumer's surpluses. In the older welfare theories, debate about value judgments disguised the importance of distinguishing between two fundamentally different concepts. One is the gain in money income attributable to a change in utility; following Hicks, we call this *equivalent variations* in money income. The other concept, which we label *compensating variations* (again following Hicks), is designed to measure offsetting changes in the budget plane that would leave the individual on the same utility level after a change in one or more of the prices faced by him. We shall show that all consumer's surpluses are variations on these two themes, and differences among them may be directly traced to what variables are being held fixed.

A welfare loss ( $WL$ ) function as constructed and extensively used, for example, by Harberger (1954, 1964) must, if it is to have the meaning generally ascribed to it, be some sort of equivalent variation in income expressed as a function of the price distortions. For example, if  $t_1, \dots, t_n$  represent excise taxes on the  $n$  commodities  $x_1, \dots, x_n$ , and if, in the absence of such taxes, the commodities were sold at prices equal to their respective marginal costs, then we would write  $WL = f(t_1, \dots, t_n)$ . It will be conclu-

\* University of Washington. I have benefited from discussions with J. Allan Hynes, Stephen N. S. Cheung and Yoram Barzel. All remaining errors are of course my own. I am indebted to Professor Yoshihiro Maruyama of Kyoto University, Japan, for providing me with a copy of M. Sono's Japanese paper on consumer's surplus, and to Kenji Kise for his help in translating that paper.

<sup>1</sup> No attempt has been made to provide an exhaustive bibliography on the consumer's surplus problem, nor will original sources necessarily be cited. Instead, the reader is referred to well-known summary treatments like Patinkin, Hicks (1959), and James Morgen (whose treatment of the problem is excellent) for references to the classical literature. A rather complete bibliography of literature on the consumer's surplus before 1960 can be found in E. J. Mishan.

<sup>2</sup> See Robert Bishop, Arnold Harberger (1964), David Schwartzman, to cite just a few examples.

sively demonstrated that such a construct is impossible, by showing the path dependence of  $WL$ , i.e., the dependence of  $WL$  upon the pattern of adjustment to the tax structure.

All the equivalent variations will be shown to be derivable from a single expression, a line integral, whose value will, in general, depend upon the particular path of price or excise tax changes. The condition under which a function such as  $WL$  is definable (i.e., the condition under which the line integral generating the equivalent variations is path-independent) will be shown to be that of constant marginal utility of money income. A precise interpretation and analysis of the "Marshallian Triangle," i.e., the area to the left of the demand curve, will be given, and its noncommensurability with the classical equivalent and compensating variations in income will be explained. After a discussion of the compensating variations, an example will be used to reveal the extent of money values that can be associated with a given utility gain experienced by a consumer when prices change. A final section relates these findings to welfare theory.

### I. The "Equivalent Variations"

In analyzing the case of a single utility-maximizing consumer the appropriate Lagrangian function is

$$L = U(x_1, \dots, x_n) + \lambda \left( M - \sum_{i=1}^n p_i x_i \right)$$

producing the usual first-order conditions

$$(1) \quad U_i - \lambda p_i = 0, \quad i = 1, \dots, n$$

$$(2) \quad M - \sum_{i=1}^n p_i x_i = 0$$

where  $U_i = \partial U / \partial x_i$ . The utility function is assumed to be strictly increasing and quasi-concave in  $x_i$ , and the Jacobian associated with equations (1) and (2) is assumed non-zero. Under these conditions these relations may be solved (in principle) for the "ordinary" (uncompensated) demand functions

$$(3) \quad x_i^* = h_i(p_1, \dots, p_n, M), \quad i = 1, \dots, n,$$

the  $*$  denoting optimal (equilibrium) values. Also produced is the equilibrium  $\lambda^*$ . The  $x_i^*$  can be substituted into  $U$ , giving

$$(4) \quad \begin{aligned} U^* &= U(x_1^*, \dots, x_n^*) \\ &= V(p_1, \dots, p_n, M) = V \end{aligned}$$

The indirect utility function,  $V$ , gives the maximum value of utility associated with a given price-income vector  $(P, M) = (p_1, \dots, p_n, M)$ .

Consider now the following *line integral*:

$$(5) \quad \Phi = \int_C \sum_{i=1}^n p_i dx_i^*$$

where  $C$  is some path of prices and income between initial and final price-income vectors  $(P^0, M^0) = (p_1^0, \dots, p_n^0, M^0)$  and  $(P^1, M^1) = (p_1^1, \dots, p_n^1, M^1)$ , respectively.<sup>3</sup>  $\Phi$  can be thought of as a generalization of the area under a demand curve, although there is more to it than that, as will be seen shortly. The adjustment process can be visualized as the gradual changing of prices and income or the changing of excise and income taxes (ignoring the destination of those tax revenues), with the consumer continuously and instantaneously adjusting to the new equilibria.<sup>4</sup> Hence the demand relations (3) hold along each point of the path of integration. To save notational clutter, the asterisks will be dropped from the variables. All values will be presumed to be in consumption equilibrium.

It is now possible to give  $\Phi$  an economic interpretation, using the first-order relations (1):

$$(6) \quad \begin{aligned} \Phi &= \int_C \sum p_i dx_i = \int_C 1/\lambda \sum U_i dx_i \\ &= \int_C 1/\lambda dU = \int_C 1/\lambda dV \end{aligned}$$

<sup>3</sup> A similar expression is used in the construction of Divisia price indices. Many of the results cited in the paper have analogous formulations in terms of such indices. See, for example, Herman Wold, William Gorman.

<sup>4</sup> A partial equilibrium approach to this problem is maintained for expositional ease. A more general approach would, of course, consider these price changes as

The Lagrange multiplier  $\lambda$  has the interpretation of the marginal utility of income. Its reciprocal,  $1/\lambda$ , can be regarded as the marginal cost of utility. It is the Lagrange multiplier associated with the cost-minimization problem,

$$\text{Minimize} \quad M = \sum_{i=1}^n x_i p_i$$

$$\text{subject to} \quad U(x_1, \dots, x_n) = U^0$$

More precisely,  $1/\lambda$  is the imputed marginal rent associated with the level of utility at a point along  $C$ . Hence the integral (6) can be regarded as the limit of a sum of marginal (dollar) rents associated with a given movement through the parameter space, and  $\Phi$  is thus the total change in dollar rent associated with a change in utility along a specified path. It must be emphasized that  $\Phi$  gives an imputed rent, or "shadow price," of utility changes and not an actual realized money income change. In general no shift in the budget plane can be uniquely associated with a given value of  $\Phi$ . The term "rent" is appropriate to the partial equilibrium framework used here in that the consumer experiences an increase in real income with the same nominal income.

The central problem of the many consumer's surpluses can now be stated in terms of a fundamental mathematical property of line integrals: Integrals such as (5) above are generally not independent of the path of integration. We must, therefore, expect to find different dollar evaluations of the same gain in utility when price and income changes follow different paths, even though the terminal price-income vectors are identical. This path-dependency is the *raison d'être* of the many consumer surpluses. The restriction to single price changes (perhaps for graphical expediency) has formerly completely hidden from view the fundamental nature of the problem.<sup>5</sup>

changes in the marginal costs of producing the commodities due to, say, technological innovation or some such resource-saving device. Even that would not be completely satisfactory, since presumably resources are expended in producing technological change.

<sup>5</sup> Although Samuelson (1947) and Harold Hotelling refer to this generalization to line integrals, they seem

One can visualize the path dependence of  $\Phi$  by noting that if, say,  $p_i$  changes, the demand curves for the other commodities begin to shift at the rate  $(\partial x_j / \partial p_i)$ ,  $j \neq i$ . If, however, some other price  $p_j$  changes, the demand for commodity  $i$  shifts at the rate  $\partial x_i / \partial p_j$ . Since these rates are not in general equal, the way in which  $p_i$  and  $p_j$  are changed—for example, first  $p_i$  then  $p_j$  or vice versa—will affect the areas under the demand curves  $\int p_i dx_i$  and hence the value of  $\Phi = \int \sum p_i dx_i$ . As it turns out, the condition  $\partial x_i / \partial p_j = \partial x_j / \partial p_i$ ,  $i, j = 1, \dots, n$  is precisely the condition for path-independence.<sup>6</sup> This will be discussed in greater detail in the next section.

Integrating by parts yields

$$\begin{aligned} \Phi &= \int_C 1/\lambda dV = \int_C \sum p_i dx_i \\ (7) \quad &= \sum p_i^1 x_i^1 - \sum p_i^0 x_i^0 - \int_C \sum x_i dp_i \end{aligned}$$

Hence the rent imputed to a movement through the parameter space has two components: the actual change in money income (independent of the path and not properly called a surplus since it involves an actual increase in money income) and a line integral which relates to price changes only and measures the consumer's imputed gain or loss due to changes in prices.

Some well-known equivalent variations in income will now be derived from the integral  $\Phi$  and analyzed.<sup>7</sup> Suppose consumer equilibrium is at  $X^0 = (x_1^0, \dots, x_n^0)$  with prices  $(P^0, M^0) = (p_1^0, \dots, p_n^0, M^0)$  on an indifference

not to regard it as crucial. Samuelson is particularly cryptic here. He states that "In general, line integrals will replace simple integrals with the path of integration of the former a matter of no consequence" (1947, p. 202n). This cannot be true as it stands. Don Patinkin, on the other hand, does mention the path dependency of consumer's surplus, although no explicit presentation of that idea is offered.

<sup>6</sup> See any advanced calculus text; for example, Angus Taylor.

<sup>7</sup> That the Hicksian consumer's surpluses are derivable from  $\Phi$  and the conditions under which the above expression is path independent, was apparently first shown by the mathematician Masuzo Sono in an article published in Japanese during World War II.

level  $U^0$ . Suppose that  $p_1$  is lowered to  $p_1^1$ , producing a new equilibrium  $X^1 = (x_1^1, \dots, x_n^1)$  on indifference level  $U^1$  where, of necessity,  $U^1 > U^0$ . What is the rent imputed to this gain in utility?

1. If  $p_1$  is lowered from  $p_1^0$  to  $p_1^1$  (for example, an excise tax is lowered on Commodity 1, ignoring the destination of those tax revenues) such that the consumer continuously and instantaneously adjusts to this price movement, then from equation (7) the imputed value of the utility gain along this path is

$$\Phi = - \int_{p_1^0}^{p_1^1} x_1 dp_1,$$

or the area to the left of the ordinary demand curve. Thus we see that this famous Marshallian area does have an unambiguous interpretation. However, *it is only one of many shadow prices* associated with the utility increase  $U^1 - U^0$ , and, indeed, only one of many shadow prices associated with this particular price change. If the path of price changes involves changes in  $p_2, \dots, p_n$ , even though the final price vector is  $(p_1^1, p_2^0, \dots, p_n^0)$ , the consumer will impute a dollar gain to the new equilibrium different from the triangular geometrical area to the left of the demand curve. An example of this situation is given in Section IV below.

2. Suppose now that, with the original prices, money income  $M$  is raised from  $M^0$  to, say,  $M^1$ , where the budget line is tangent to  $U^1$  at the old prices. Then, clearly,  $\Phi$  equals this change in money income. (The line integral  $-\int \sum x_i dp_i$  is zero since prices are constant.) To make the terminal point the same as in paragraph 1 above, after the budget shift decrease  $M$  and all prices by the same proportion, returning  $M$  to  $M^0$ , then change relative prices so as to move the consumer along  $U^1$  to the point  $X^1$  at price-income vector:  $(p_1^1, p_2^0, \dots, p_n^0, M^0)$ . Since  $U$  is constant,  $dU = 0$ ; hence  $\Phi$  is zero for these latter operations. Thus  $\Phi$  gives the classical equivalent variation in income associated with a rise in a price as defined by Hicks (1959). The classical equivalent variation in income for a price decrease (see Hicks (1959)) is yielded by a

similar procedure whereby relative prices are changed first so as to move the consumer along  $U^0$  until the prices are proportional to the new prices  $(p_1^1, p_2^0, \dots, p_n^0)$  and then  $M$  is increased.<sup>8</sup> There is, of course, no reason why this second equivalent variation should equal the first one; indeed, they generally differ. Both represent imputed dollar rents associated with the utility gain  $U^1 - U^0$ , but for different paths of achieving that gain.

Clearly, an infinite number of consumer's surpluses can be generated, most of them uninteresting. Relative and absolute prices can be changed so as to hold  $x_1$  fixed until  $U^1$  is reached and then changed relatively so as to move the consumer along  $U^1$  to  $X^1$ . Alternatively,  $x_2, \dots, x_n$  can be held constant. These paths generate values of  $\Phi$  which correspond to (but are not exactly equal to) Marshall's surpluses in which quantities are held constant. There is no way in general to relate the magnitude of these surpluses to other values of  $\Phi$ .

It is necessary now to show that the equivalent variations described in paragraph 2 above cannot meaningfully be compared with the area to the left of the demand curve. Since  $1/\lambda = M / \sum U_i x_i$  and a proportional change in prices and income leaves the denominator unchanged, it is clear that  $1/\lambda$  is homogeneous of degree 1 in prices and income (Samuelson (1942)). Therefore, the integral  $\Phi = \int 1/\lambda dU$  is not invariant with respect to proportional changes in prices and income. With nonnormalized prices (i.e., with  $M$  allowed to vary) no unique value of  $\Phi$  can be associated with a path in the commodity space—say, the price-consumption or income-consumption paths. Consider now the following "equivalent variation": Starting at point  $X^0$  at  $(P^0, M^0)$ , hold  $M$  constant at  $M^0$  and slowly deflate all prices proportionally until the budget line is tangent to  $U^1$ . Evaluated on this path,  $\Phi$  is less than the corresponding equivalent variation in paragraph 2 above when  $M$  was increased. The reason is that at every point along the income-consumption path between  $U^0$  and  $U^1$  beyond

<sup>8</sup> This is equivalent to the derivation of Laspeyres and Paasche indices from a Divisia index. See, for example, Wold.

$X^0$  at the original price ratios,  $1/\lambda$  for the price-deflating case is less than  $1/\lambda$  for the  $M$  increasing case. This is easily seen from the above definition of  $1/\lambda$ . While the same bundle may be purchased, an extra dollar represents more purchasing power if prices have been deflated than if  $M$  had increased. Hence the marginal utility of money income is greater, or the imputed rent to an extra utile is less when prices have deflated rather than when income has increased. Clearly, then, it makes no sense to compare a consumer's surplus where  $M$  has been held fixed (say, the area to the left of the demand curve as in paragraph 1 above) with one in which  $M$  has varied. The dollar amounts in the two cases are not commensurable, since the imputed rents are based on differing dollar bases.

It is therefore apparent that the values of  $\Phi$  derived from the removal of excise taxes or tariffs or the removal of monopolistic pricing policies can in no way be used as a conceptual basis in constructing a welfare loss function for the community or in applying a compensation principle. This is so for two major reasons. First, the very value of  $\Phi$  is in general dependent upon the adjustment path from one equilibrium to the next; no equivalent variation is a function solely of the terminal price-quantity coordinates. Second, the value of equivalent variations derived when money income is held constant is not commensurate with a value derived when money income has varied, as in the case of an actual income transfer.

## II. The Constancy of the Marginal Utility of Income

We now inquire as to the conditions under which these surpluses have a common value. Since  $\lambda$  varies when all prices and income change in the same proportion, this question is meaningful only if  $M$  is held constant. We are then seeking to find out when

$$\Phi = \int \sum p_i dx_i = - \int \sum x_i dp_i$$

is independent of the path of integration. This integrability holds if  $\partial x_i / \partial p_j = \partial x_j / \partial p_i$ ,

$i, j = 1, \dots, n$ . As is well known (see Samuelson (1942, 1947); Laurence Lau), this condition implies that the utility function is homothetic (i.e., the income-consumption paths are rays emanating from the origin) or that the income elasticities are all unity.<sup>9</sup>

Geometrically, the relation between the exactness of  $\Phi$  and homotheticity is as follows. Suppose the utility function is homothetic and the consumer is at point  $X^0$  at prices  $P^0$  on utility level  $U^0$ . Now decrease prices or increase  $M$  by some proportion. The consumer will now attain some higher utility level  $U^1$ . Suppose now that he had started at some other point along  $U^0$ , say  $X^{0'}$  at prices  $P^{0'}$ . Suppose these prices are lowered, or  $M$  increased, by the same proportion as before. Then, since the indifference curves are radial blow-ups of each other, the consumer must achieve a point again on  $U^1$ . Thus in this case of homothetic utility functions there is a unique correspondence between utility gains and proportions of income gains, no matter where the consumer starts on  $U^0$ ; i.e., no matter what price vector he faces initially. This suggests that for homothetic indifference maps,  $\lambda$ , the marginal utility of income, is independent of prices. Indeed, this is the case. As Lau showed,  $1/\lambda = 1/\lambda(U) = 1/\lambda(V)$  if, and only if, the utility function  $U$  (and hence the indirect utility function  $V$ ) is homothetic. If  $1/\lambda = 1/\lambda(V)$  then

$$\Phi = \int_{P^0}^{P^1} 1/\lambda(V) dV$$

is an ordinary definite integral, dependent only on the initial and final utility levels and not on the path of adjustment.

This analysis can be approached from another direction. From equation (7),  $\lambda$  is seen to be an integrating factor for  $d\Phi = \sum x_i dp_i$ .

<sup>9</sup> Using the Slutsky equation

$$\frac{\partial x_i}{\partial p_j} = \left( \frac{\partial x_i}{\partial p_j} \right)_U - x_j \frac{\partial x_i}{\partial M},$$

and noting that

$$\left( \frac{\partial x_i}{\partial p_j} \right)_U = \left( \frac{\partial x_j}{\partial p_i} \right)_U$$

always, it immediately follows that all the income elasticities must be unitary.

In other words, the "reciprocity" relation  $\partial(\lambda x_i)/\partial p_j = \partial(\lambda x_j)/\partial p_i$ ,  $i, j = 1, \dots, n$  holds for all utility functions; hence  $\lambda d\Phi$  is always an exact differential.<sup>10</sup> If  $\lambda$  is constant, i.e., independent of prices, it can be removed from the integral sign, yielding

$$\begin{aligned} [\Phi(P^1) - \Phi(P^0)] &= (1/\lambda) \int \sum \lambda x_i dp_i \\ (8) \qquad \qquad \qquad &= \left(1/\lambda \int dV\right) \\ &= (1/\lambda)(V^1 - V^0) \end{aligned}$$

Hence, in this case the imputed dollar gain is proportional to the utility gain.

To sum up, all the equivalent variations in income are equal when, and only when, the utility function is homothetic. Empirically, alas, this case is unimportant;<sup>11</sup> but in it the line integral generating the shadow prices of utility changes is independent of the adjustment path. When the utility function is not homothetic, literally an infinity of imputed rents can be associated with any given utility change.

### III. The Compensating Variations

We now come to the second major class of "consumer's surpluses," the compensating

<sup>10</sup> This can also be proved by direct differentiation:

$$\begin{aligned} \frac{\partial(\lambda x_i)}{\partial p_j} &= \frac{1}{D} (\lambda^2 D_{ij} + x_j D_{n+1,i} + x_i D_{n+1,j} \\ &\quad + x_i x_j D_{n+1,n+1}) = \frac{\partial(\lambda x_j)}{\partial p_i}, \end{aligned}$$

where  $D$  is the determinant of the bordered Hessian matrix associated with the first-order equations (1) and (2), with  $D_{ij}$  the cofactor of the element in the  $i$ th row and  $j$ th column.

Using an example of Samuelson's (1942), if  $U$  is the particular homothetic function  $U = \log W$ , where  $W$  is homogeneous of degree 1, then

$$\lambda = \sum U_i x_i / M = 1/M \cdot 1/W \sum W_i x_i = 1/M$$

Hence  $\lambda$  is constant with respect to prices. Since all homothetic utility functions can be expressed as a monotonic transformation of a function of the type above ( $U = \log W$ ), say,  $Z = F(U)$ ,  $\lambda_Z = F' \cdot \lambda_U$  is similarly independent of prices.

<sup>11</sup> There are, however, important special cases where homotheticity may hold—specifically, in intertemporal analysis. See, for example, Milton Friedman's analysis of the consumption function.

variations, defined as the change in money income needed to exactly offset the gain (loss) in utility due to a fall (rise) in one or more prices. In the context of the present analysis the concept is an attempt to avoid the path-depending problems associated with the equivalent variations. Differentiating the budget constraint totally,

$$dM = \sum x_i dp_i + \sum p_i dx_i$$

Since the individual remains on the same indifference level,

$$dU = dV = \lambda \sum p_i dx_i = 0$$

Hence,

$$(10) \quad M^0 - M^1 = - \int_{p_0}^{p_1} \sum x_i dp_i$$

independently of the path of integration. Alternatively, since  $U$  is constant, the symmetry of the substitution matrix yields  $\partial x_i / \partial p_j = \partial x_j / \partial p_i$ ; hence when income-compensated demands are used,  $\sum x_i dp_i$  is an exact differential. When only one price changes, say  $p_1$ , then the compensating change in income equals the area to the left of the compensated demand curve, or

$$- \int_{p_0}^{p_1} x_1 dp_1$$

This measure is not, in general, invariant with respect to the utility level the consumer is held to; only in the case of vertically parallel indifference curves will this special property occur.<sup>12</sup>

The significance of the path independence of the compensating variations is that it is possible to associate a unique money income change with any change in prices, as long as

$$\begin{aligned} &^{12} \frac{\partial(M^0 - M^1)}{\partial U^0} \\ &= - \int_{p_1^0}^{p_1^1} \frac{\partial x_1}{\partial U^0} dp_1 \gtrless 0 \quad \text{as} \quad \frac{\partial x_1}{\partial U^0} \gtrless 0 \end{aligned}$$

i.e., as  $x_1$  is a superior or inferior good. The intermediate case (invariance with respect to the level of utility of the compensating variation) is a vertical income-consumption path. See, for example, Patinkin.

the consumer remains on the same indifference level. The only data needed to calculate the compensating changes in income are the terminal price vectors. No attention need be given to the adjustment path; hence, it is possible to meaningfully formulate questions such as, "How much will an individual pay for the privilege of purchasing a commodity at a lower price?" or "How much compensation is needed for an individual to voluntarily accept a higher price?" or "How much in addition to what a consumer now pays for the amount  $x$  of good  $X$  would the consumer be willing to pay for  $x$  rather than go without any  $X$ ?" These questions, which relate to changes in money income needed to hold utility constant, have meaningful, unique answers because of the path-independence property of compensating variations.

This path independence, though not generally recognized as such, has led economists to attempt to substitute the compensating variations for the equivalent variations in income, as a welfare measure. These efforts are futile, as we now demonstrate for the particular case of a fall in one price.

The ordinary demand curve lies to the right of the compensated demand curve in the case of a "normal" good as price is lowered; hence the dollar gain imputed to a lowering of, say,  $p_1$

$$\Phi = - \int_{p_1^0}^{p_1^1} x_1 dp_1$$

is greater than the corresponding compensating variation, or vice versa in the case of an inferior good. However, in the light of the discussion relating to the meaning of equivalent variations when money income changed, it is apparent that these equivalent and compensating variations cannot meaningfully be compared. As Hicks (1959) showed, the compensating variation above, where  $U = U^0$ , equals the equivalent variation for a fall in price, as described in Section I, paragraph 2 above. As shown there, this equivalent variation is not commensurate with the area to the left of the uncompensated demand curve. It is true that for "small" variations in price (i.e., if the price change causes small changes

in utility) the two areas approximate each other. But equivalent variations and compensating variations are clearly two different concepts. The former imputes a dollar evaluation to a change in utility levels for a particular path of price changes, while the latter derives dollar values necessary to hold utility constant when prices change, over any path of adjustment. There is no reason to expect these values to be comparable in any way.

Consider again the area to the left of the uncompensated demand curve. Graphical analyses of this area made by Harberger (1964) and by David Winch agree with the approach presented in this paper insofar as explicit attention is given to the adjustment path. Winch correctly derives the "Marshallian Triangle" as the limit of a sum of compensating variations in income that a consumer would be willing to pay for the privilege of consuming at a slightly lower price, always assuming, however, that he never actually pays those amounts. Winch did not show, however, that all of the equivalent variations can be so derived. If a different adjustment path of prices were specified, the limit of the sum of the compensating variations the consumer would be *willing* to pay (not actually pay) along this new path would, in general, differ from the original sum, even though the initial and final price vectors were not changed. This crucial point alone invalidates the assertion that the Marshallian area is *the* consumer's gain.

If the consumer actually makes payment at each point along the adjustment path he would no longer be on his ordinary demand curve but rather on his compensated curve. In noting this aspect of the Marshallian area, Winch described it as the "consumer's gain." A characteristic of this gain, or shadow price as it is called here, is that it can be captured only by the consumer. A perfectly discriminating monopolist having sufficient information could capture an amount equal to the area to the left of the compensated demand curve (which for noninferior goods is less than the consumer's gain). This case, in which any attempt to manipulate the experiment fundamentally changes the phenomenon being observed, further indicates the

futility of using equivalent variations as welfare measures.

Harberger (1964) is less precise about the areas involved, and eventually he returns to the compensated curves. His analysis, dealing explicitly with the shifting of demand curves for the commodities whose prices have not changed, is precisely the evaluation of the line integral

$$\Phi = \int \sum p_i dx_i$$

for a single price change, his rectangles being  $p_i^0(x_i^0 - x_i^1)$ ,  $i=2, \dots, n$ .<sup>13</sup> However, both Harberger (1964) and Winch ignored the natural generalization of their analyses presented in this paper and therefore did not explore the path-dependency aspects of their results.<sup>14</sup> Their analyses mistakenly imply that the derived welfare gains or losses are unique with respect to the terminal price coordinates. As we have shown, and will perhaps clarify in an example below, an infinity of such shadow gains is associated with the simple act of lowering even one price.

#### IV. Extreme Values of $\Phi$ : An Important Example

Returning now to the general unrestricted case of stable ordinary demand curves, we inquire whether such a path may exist between  $(P^0, M^0)$  and  $(P^1, M^0)$  that the line integral  $\Phi$  has an interior extremum. This can be formulated as:

$$(12) \quad \max - \int_{p^0}^{p^1} \sum x_i dp_i \quad \text{or} \\ \min \int_{p^0}^{p^1} \sum x_i dp_i$$

The necessary conditions for these two problems are the same; we consider only the maximum case. Parameterizing this as  $p_i = p_i(t)$ ,  $p_i^0 = p_i(t^0)$ ,  $p_i^1 = p_i(t^1)$ , one gets

<sup>13</sup> Harberger deals with "tax revenue constant" demand curves, hence his resulting area does not correspond exactly to that presented in this paper.

<sup>14</sup> In an earlier article, Harberger (1954) mentions the path dependency of his results but chooses to assume away the problem.

$$(13) \quad \max - \int_{T^0}^{T^1} (\sum x_i \dot{p}_i) dt$$

Since this is a fixed end-point problem, the Euler-Lagrange conditions of control theory can be applied. They yield the matrix equation

$$(14) \quad (x_{ij} - x_{ji})(\dot{p}_i) = (0),$$

where

$$x_{ij} = \frac{\partial x_i}{\partial p_j}$$

Nontrivial solutions can appear only if the matrix  $(x_{ij} - x_{ji})$  is singular. Since this is not the case in general,  $\Phi$  will not have an interior maximum or minimum. A special case where (14) is satisfied is when  $\partial x_i / \partial p_j = \partial x_j / \partial p_i$ . Under these conditions, all paths generate the same value of  $\Phi$ , hence a trivial extremum. The linearity in  $\dot{p}_i$  of the integrand in (13) is clearly the reason why no interior extremum can exist.

As an example of the possibilities for identifying money income changes with given utility changes, consider the utility function  $U = (\log x_1) + x_2$ .<sup>15</sup> (See Figure 1.) Since  $M$  is

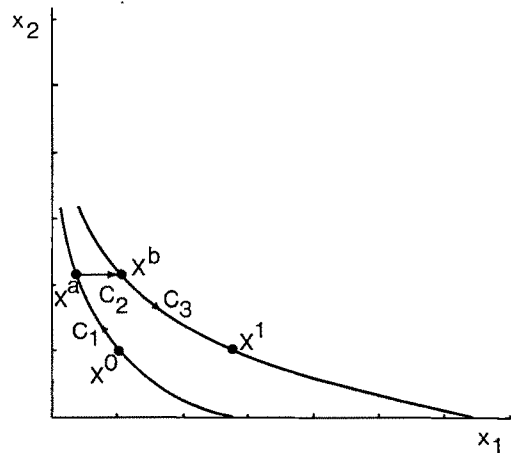


FIGURE 1

<sup>15</sup> The indifference curves here are vertically parallel, and the price consumption path for  $x_1$  is horizontal. These properties make the arithmetic easier but in no way destroy the meaning of the results, since these properties are irrelevant to equivalent variations.

to be held constant, let  $M=1$ . The associated demand curves are

$$x_1 = p_2/p_1, x_2 = 1/p_2 - 1, p_2 < 1$$

Let the consumer be initially at  $X^0 = (1, 1)$  at prices  $P^0 = (1/2, 1/2)$  with  $U^0 = 1$ . Suppose  $p_1$  is lowered to  $1/2e$ , yielding the terminal point  $X^1 = (e, 1)$ ,  $P^1 = (1/2e, 1/2)$ ;  $U^1 = 2$ . What dollar values can the consumer impute to this change in utility?

Let us evaluate  $\Phi = \int_C 1/\lambda dV$  over the following path. Starting at  $P^0$ , change relative prices so as to move the consumer along  $U^0$  to some point  $X^a$  at  $(p_1^a, p_2^a)$ . (Call this branch  $C_1$ .) Now lower  $p_1$ , moving the consumer out to  $X^b$  on  $U^1$  at prices  $(p_1^b, p_2^a)$ . (Call this path  $C_2$ .) Now change relative prices to move the consumer along  $U^1$  to  $X^1$  at  $P^1$  above (path  $C_3$ ). Along  $C_1$  and  $C_3$   $\Phi$  is zero. Along  $C_2$ , since  $p_2$  is constant,

$$\begin{aligned} \Phi &= - \int_{p_1^a}^{p_1^b} x_1 dp_1 = - \int_{p_1^a}^{p_1^b} \frac{p_2^a}{p_1} dp_1 \\ (12) \quad &= -p_2^a \log p_1 \Big|_{p_1^a}^{p_1^b} \\ &= p_2^a (\log p_1^a - \log p_2^a) \end{aligned}$$

The indirect utility function is  $V = \log p_2 - \log p_1 + 1/p_2 - 1$ . Hence, the change in utility in moving from  $X^a$  to  $X^b$  is

$$\begin{aligned} V^b - V^a &= \left( \log p_2^a - \log p_1^b + \frac{1}{p_2^a} - 1 \right) \\ (13) \quad &- \left( \log p_2^a - \log p_1^a + \frac{1}{p_2^a} - 1 \right) \\ &= \log p_1^a - \log p_1^b \end{aligned}$$

Substituting this into  $\Phi$  yields

$$(14) \quad \Phi = p_2^a (V^b - V^a)$$

In this example  $V^b - V^a = U^1 - U^0 = 1$ , and  $0 < p_2^a < 1$ . Hence the imputed value of the one unit gain in utility can be any real num-

ber between zero and one! No reason appears why an example could not be constructed yielding any real number as an imputed value of a finite utility gain. The upper bound on  $\Phi$  in this example is produced by the intersection of the indifference curves with the horizontal axis.

#### V. Concluding Remarks: Are the Equivalent and Compensating Variations in Income of Any Use in Economic Analysis?

The above analysis and example demonstrate the arbitrary nature of the assignment of dollar values to utility changes. If the price change is abrupt and discontinuous in nature, then simply no dollar value can be associated with the resulting equilibrium change.

Suppose a dam is constructed which lowers the marginal costs of electricity, local recreation, and irrigation. In evaluating what benefit consumers would place on this project, some assumption must be made about the price adjustment path. Curiously paradoxical is the fact that a given consumer will impute a different shadow price to the project if first the price of electricity is lowered, then that of irrigation, and then of recreation, than if prices were lowered in some other sequence; yet the nagging truth remains that nothing can be done to correct such an inconsistency. It is therefore time, at long last, for economists to abandon the term "welfare loss" in their discussions of monopolies, tariffs, etc. If the term were merely ambiguous, the analysis could be narrowed and at some cost the term could be meaningfully applied. However, the phrase simply has no meaning at all. It is impossible to construct the function depicted in the introduction to this paper which designates welfare loss as a function of a set of price distortions.

Where, then, are we to turn? Are economists doomed to making strictly ordinal comparisons of alternative equilibria? Not quite, I think. The compensating variations in income are not path dependent and are, hence, well defined. However, to merely substitute the compensating variations for the equivalent variations as a measure of welfare loss or

gain (as Harberger (1954, 1964) and others do) is to use the inappropriate to measure the undefinable.<sup>16</sup> The following alternative, however, is within the realm of positive economics: statistically estimated ordinary demand curves, together, perhaps, with estimates of income elasticities can be used to obtain approximations to the compensating variations in income. When a public-policy decision implies that some will be net losers, the compensating variations give well-defined measures of how much money income would have to be transferred to those individuals to restore them to a just-as-preferred position. Similarly, estimates can be derived as to the amounts which could be extracted from the net gainers and still leave them at just-as-preferred positions. Data of this sort can be provided; however, this is all that the economist, *qua* economist can provide.<sup>17</sup>

It is now generally recognized that where a public policy will result in both losers and gainers, value judgments are involved in the decisions whether to implement the policy at all and, if so, whether to make compensations. These decisions are necessarily made via the political process; however, by providing knowledge of the relevant compensating variations associated with policy decisions, economists may contribute information useful in the formulating of means of payment and the policing of benefits.

Scientific objectivity requires that areas to the left of empirical demand curves be regarded as estimates of compensating variations in money income and not some vague idea of consumers' gains from utility changes. In this context, consumer's surplus may be a

useful policy tool. It would be nice if there were always a unique one-to-one correspondence between utility changes and money-income changes. Much of welfare economics would be drastically simplified. However, it simply is not so.

## REFERENCES

- R. C. Bishop, "The Effects of Specific and Ad Valorem Taxes," *Quart. J. Econ.*, May 1968, 82, 198-218.
- M. Friedman, *A Theory of the Consumption Function*, Princeton 1957.
- W. M. Gorman, "Notes on Divisia Indices," unpublished.
- A. C. Harberger, "Monopoly and Resource Allocation," *Amer. Econ. Rev. Proc.*, May 1954, 44, 77-87.
- , "Taxation, Resource Allocation, and Welfare," in *The Role of Direct and Indirect Taxes in the Federal Revenue System*, Princeton 1964.
- J. R. Hicks, *A Revision of Demand Theory*, Oxford 1959.
- , *Value and Capital*, Oxford 1946.
- H. Hotelling, "The General Welfare in Relation to Problems of Taxation and of Railway and Utility Rates," *Econometrica*, 1938, 6, 242-269; reprinted in *Readings in Welfare Economics*, Homewood 1969.
- H. S. Houthakker, "Additive Preferences," *Econometrica*, Apr. 1960, 28, 244-57.
- , "A Note on Self Dual Preferences," *Econometrica*, Oct. 1965, 33, 797-801.
- L. J. Lau, "Duality and the Structure of Utility Functions," *J. Econ. Theor.*, Dec. 1969, 1, 374-95.
- E. J. Mishan, "A Survey of Welfare Economics, 1939-59," *Econ. J.*, June 1960, 70, 197-256.
- J. N. Morgan, "The Measurement of Gains and Losses," *Quart. J. Econ.*, Feb. 1948, 62, 287-308.
- D. Patinkin, "Demand Curves and Consumer's Surplus," in C. F. Christ et al., eds., *Measurement in Economics*, Stanford 1963.
- P. A. Samuelson, "The Constancy of the Marginal Utility of Income," in O. Lange et al., eds., *Studies in Mathematical Economics and Econometrics in Memory of Henry Schultz*,

<sup>16</sup> It seems strange, at best, to use as a measure of consumer's benefits a construct which explicitly assumes that individuals remain at the same level of utility.

<sup>17</sup> The above analysis should not be construed to mean that the concept of a consumer's surplus as a compensating variation is not a useful construction in positive economic analysis. The test for its usefulness lies in the refutable hypothesis it generates. I feel that such a consumer's surplus concept is probably of substantial use in the analysis of such problems as monopolistic price discrimination, price controls, and theories of political behavior, to mention just a few.

Chicago 1942; reprinted in *The Collected Scientific Papers of Paul A. Samuelson*, Cambridge 1965.

———, *Foundations of Economic Analysis*, Cambridge 1947.

———, "Using Full Duality to Show that Simultaneously Additive Direct and Indirect Utilities Implies Unitary Price Elasticity of Demand," *Econometrica*, Oct. 1965, 33, 781-96.

D. Schwartzman, "The Burden of Monopoly," *J. Polit. Econ.*, Dec. 1960, 68, 627-30.

M. Sono, "Relative Effects of Output from Choice Theory," *Keizai Ronso*, 1943, 6, 57.

A. E. Taylor, *Advanced Calculus*, New York 1955.

D. M. Winch, "Consumers Surplus and the Compensation Principle," *Amer. Econ. Rev.*, June 1965, 55, 395-423.

H. Wold, *Demand Analysis*, New York 1953.

# Fiscal and Monetary Policy Reconsidered: Comment

By JACK VERNON\*

Robert Eisner, in a recent issue of this *Review*, applies the permanent income hypothesis to the 1968-69 tax surcharge experience and makes the plausible and important point that we should not have expected that the surcharge would have a significant impact on the inflationary pressures of the period. He argues that since the surcharge was announced to be temporary, and scheduled to expire in slightly more than a year, it was not regarded by spending units as having an important impact on permanent disposable income, and therefore did not affect expenditures significantly.

A problem arises, however, with Eisner's analysis of the mechanics of the surcharge impact. He argues that the impact included 1) the initial, tax-generated impact on disposable income and expenditure and 2) secondary or subsequent rounds of impacts on expenditure as secondary income recipients exercised their propensities to consume (see pp. 899-900). He computes the depressant effect of the surcharge on money income as

$$\begin{aligned} (Y_2 - Y_1) &= \frac{b_i(t_2 - t_1)}{1 - b_s(1 - t_2)} Y_1 \\ (1) \quad &= \frac{.1(.198 - .18)}{1 - .5(1 - .198)} Y_1 \\ &= .003 Y_1, \end{aligned}$$

or somewhat less than \$3 billion, given the level of current dollar income which prevailed at the time.<sup>1</sup> The parameters  $t_1$  and  $t_2$  are the tax rates and  $Y_1$  and  $Y_2$  the income equilibria, before and after the tax rate changes, respectively, and  $b_i$  is the marginal propensity to consume from the initial, tax-generated change in disposable income, while  $b_s$  is the marginal propensity to consume from

secondary rounds of changes in disposable income. He argues that whereas both  $b_i$  and  $b_s$  are smaller than the marginal propensity to consume from permanent disposable income,  $b_s$  is larger than  $b_i$  because secondary income recipients are less able to distinguish tax-induced changes in income from variations in income arising in other factors.

The purpose of this comment is to point out that Eisner's equation (1) representation and its supporting discussion imply an expenditure function with money illusion, and that this money illusion is the source of the significant role he obtains for  $b_s$ . In the traditional, no-money illusion representation of the impact of a tax surcharge on inflationary pressures, the role of  $b_s$  is absent, or at least much reduced.

The argument is as follows. If the expenditure function is specified without money illusion, with  $b$  being the marginal propensity to make real consumption expenditures from real disposable income, secondary rounds of impacts of changes in real income and real disposable income on expenditure are not involved significantly in the impact of a tax surcharge on inflationary pressures, because real income does not change to any great extent as inflationary pressures are damped. Equation (1) has general applicability in the no-money illusion model only for "unemployment" cases, where changes in money income primarily are changes in real income.

Figures 1 and 2 illustrate these points, exaggerating them somewhat due to the simplifying assumption that responses of real income and prices to expenditure break sharply at full employment, with equilibrium prices changing and real income constant in excess demand cases, and real income adjusting to expenditure with prices constant where expenditure is insufficient to support full employment. On this assumption, the impact of a tax surcharge on inflationary pressures is entirely an impact on price equilibrium.

Figure 1 presents the Eisner representation

\* Associate professor of economics, University of Florida.

<sup>1</sup> Eisner bases the estimates for  $b_s$ ,  $t_1$ , and  $t_2$  on equations (A.1) to (A.5) and on (A.34) in Gary Fromm and Paul Taubman, pp. 126-27 and p. 133.

and demonstrates that equation (1) derives from a model featuring an expenditure function with money illusion. Figure 2 presents the traditional representation and demonstrates that Eisner's  $b_s$  is not involved when money illusion is absent.

### I. The Money Illusion Model

For Figure 1, the model is

$$(2) \quad E = E_0 + bYd$$

$$(3) \quad Yd = Y - T$$

$$(4) \quad T = tY$$

$$(5) \quad Y = E$$

where  $E$ ,  $Yd$ ,  $Y$ , and  $T$  are aggregate money expenditure, money disposable income, money income, and money taxes, respectively, and  $E_0$ ,  $t$ , and  $b$ —the marginal propensity to consume and spend from money disposable income—are constants. Money illusion is involved in equation (2), since real expenditure will vary when prices alone vary. The money illusion results from  $E_0$ , which is fixed in money terms.

Equilibrium money income in this model is

$$(6) \quad Y = \frac{E_0}{1 - b(1 - t)}$$

whether this value exceeds, equals, or falls short of  $Y_f$ , with  $Y_f$  being the money income and expenditure which purchases the full employment product at existing prices.

In Figure 1, we have the inflation case, since the equilibrium money income with  $t_1$ , the initial tax rate, is  $Y_1$ , which exceeds  $Y_f$ . The money income gap is  $(Y_1 - Y_f)$ , which is entirely an inflationary gap, since full employment real income is by assumption fixed, i.e.,

$$(7) \quad (Y_1 - Y_f) = (P_1 - P_0)Y_f^*$$

where  $Y_f^*$  is full employment real income and  $P_1$  and  $P_0$  are the equilibrium and existing price levels, respectively, with  $P$  being 1.0 in the base year.

The impact on the money income gap of an increase in tax rate to  $t_2$  is

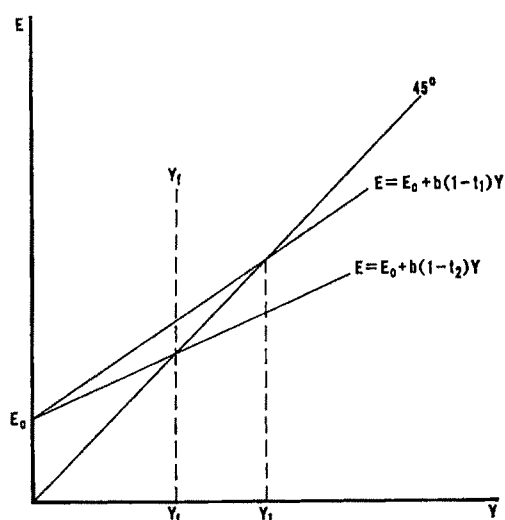


FIGURE 1

$$(8) \quad \begin{aligned} -(Y_1 - Y_f) &= (\partial Y / \partial t)(t_2 - t_1) \\ &= \frac{-b_i Y_1(t_2 - t_1)}{1 - b_s(1 - t_2)} \end{aligned}$$

where  $t_2$  is by assumption the tax rate which just closes the gap, leaving  $P_0$  the new equilibrium price level. The  $b$  in the numerator is Eisner's  $b_i$ , with  $[-Y_1(t_2 - t_1)]$  and  $[-b_i Y_1(t_2 - t_1)]$  being the initial, tax-generated impact on money disposable income and money expenditure, respectively. Secondary rounds of impacts on money income and expenditure sum to equation (8) less  $[-b_i Y_1(t_2 - t_1)]$ , the  $b$  of the denominator of equation (8) being Eisner's  $b_s$ . Equation (8) therefore is the same as equation (1). The secondary rounds of changes in money disposable income are entirely changes in prices, since real income is by assumption constant, but these price induced changes in money disposable income induce changes in money expenditure through  $b_s$  because money expenditure is a function of money disposable income.

### II. The No-Money Illusion Model

The model for Figure 2 is

$$(9) \quad E = E_0 + bYd - vr$$

$$(10) \quad Yd = Y - T$$

- (11)  $T = tY$   
 (12)  $Ms = Ms_0/P$   
 (13)  $Md = mY - Lr$   
 (14)  $Y = E$   
 (15)  $Ms = Md$

where  $E$ ,  $Y_d$ ,  $Y$ ,  $T$ ,  $Ms$ , and  $Md$  are aggregate expenditure, disposable income, income, taxes, money supply, and money demand, respectively, all real variables;  $P$  is the price level;  $r$  is the rate of interest;  $E_0$  is the expenditure constant; and  $Ms_0$  is the nominal money supply. The slope parameters  $b$ ,  $t$ ,  $v$ ,  $m$ , and  $L$  are constants. The linear form is used to facilitate comparison with Eisner's equation (1) representation.

This model is free from money illusion in the sense that real expenditure, real disposable income, real taxes, and real money balances demanded will not vary with equiproportionate changes in prices, money income, and the money stock.<sup>2</sup> The expenditure function in particular is not a source of money illusion as it is in Eisner's model, since real expenditure will not vary when prices vary with real disposable income and interest rate constant.<sup>3</sup>

We have the inflationary gap case in Figure 2, since the intersection of the  $LM$  schedule for existing prices,  $LM_{P_0}$ , and the  $IS$  schedule at the initial tax rate,  $IS_{t_1}$ , is beyond  $Y_f$ , the full employment real income.

<sup>2</sup> The tax function is an abstraction, of course. Real taxes and real disposable income in fact do vary somewhat when money income varies with real income constant since tax rates apply to money income and the tax function is not proportionate but progressive with a constant relatively fixed in dollar terms. This would provide some role for  $b$ , even where the expenditure function itself is not a source of money illusion and real income does not change with the surcharge impact. The tax function is not a source of money illusion in Eisner's model, however, since his tax function is proportionate.

<sup>3</sup> The criteria for money illusion employed here is consistent with that offered by Don Patinkin, pp. 22-23.

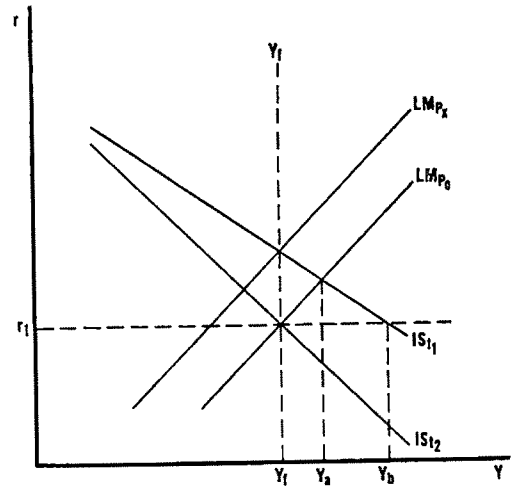


FIGURE 2

The inflationary gap is  $(P_x - P_0)$ , or  $(P_x - P_0)Y_f$  in money income terms, where  $P_x$  and  $P_0$  are the equilibrium and existing price levels, respectively, and condition (16) holds.

The impact of an increase in  $t$  to  $t_2$  on  $(P_x - P_0)$  is reflected in Figure 2 in the shift of the  $IS$  schedule to  $IS_2$ . Again the presentation is such that  $t_2$  is just sufficient to eliminate the inflationary pressures, leaving  $P_0$  the equilibrium price level.

The differential

$$(17) \quad dP = (\partial P / \partial t) dt \\ = [(P_x^2 / Ms_0)(L/v)(-bY_f)] dt$$

is useful in indicating the mechanics of the surcharge impact on price equilibrium, though it can be used as an approximation of the impact itself only where the change in  $t$  is small. The portion  $[-bY_f dt]$  is the impact on the expenditure gap at  $P_0$  and  $t_1$ , i.e., the impact on

$$(18) \quad [E = E_0 + b(1 - t_1)Y_f - vr_1] - Y_f,$$

since

$$(19) \quad dE = (\partial E / \partial t) dt = -bY_f dt,$$

$$(16) \quad \left[ P_x = \frac{Ms_0}{Y_f \{ (L/v)[1 - b(1 - t_1)] + m \} - (L/v)E_0} \right] > P_0$$

or, in Eisner's terminology,  $[-bY_f \Delta t]$  is the initial, tax-generated impact on expenditure, with the  $b$  being  $b_i$ . Eisner's  $b_s$ , representing secondary rounds of impact on expenditure as secondary income recipients exercise their propensities to consume, does not enter into the impact on the expenditure gap because real income does not change.<sup>4</sup> The  $b$  parameter also enters into equation (17) in that it influences  $P_x$ , the initial equilibrium price level, of course, but this clearly is not Eisner's secondary income effect involvement.  $P_x$  is involved in the impact on the price gap because the amount by which the price level must change to effect a particular reduction in real money balances varies with the initial price level and money stock.

### III. Money Illusion with Money Market Influences

While the primary purpose of this comment has been to point out that money illusion is implied in Eisner's equation (1) and that the role of  $b_s$  is absent or much reduced when money illusion is not present, Eisner's representation also differs from the traditional one in its neglect of money market influences. Even if the money illusion assumption is appropriate, which is not out of the question,<sup>5</sup> equation (1) still should be criticized for neglecting money market influences, since these influences weaken the surcharge impact. For example, if equations (9) through (15), excluding (12), and the horizontal axis of Figure 2 are reinterpreted so that  $E$ ,  $Y_d$ ,  $T$ ,  $M_d$ ,  $M_s$ , and  $Y$  are money variables rather than real variables, and equation (12) is replaced by  $M_s = M_{s0}$ ,<sup>6</sup> the equilibrium money income is as  $Y_a$  in Figure 2, where

$$-(Y_b - Y_f) = \frac{-b_i Y_b (t_2 - t_1)}{1 - b_s (1 - t_2)},$$

but  $[-b_i Y_f (t_2 - t_1)]$

<sup>5</sup> See William Branson and Alvin Klevorick.

<sup>6</sup> For analytical convenience, these assumptions specify both the money demand and the expenditure function as sources of money illusion.

$$(20) \quad Y_a = \frac{LE_0 + vMs_0}{vm + L[1 - b(1 - t)]}$$

and the impact of the increase in tax rate to  $t_2$  is

$$(21) \quad -(Y_a - Y_f) = (\partial Y / \partial t)(t_2 - t_1) \\ = \frac{-b_i Y_a (t_2 - t_1)}{(vm)/L + [1 - b_s(1 - t_2)]}$$

rather than as in equation (1).

While Eisner does not give the reason for his neglect of money market influences in equation (1), such neglect is consistent with his view, expressed later (1969), p. 904, that the interest responsiveness of money demand is in effect extremely large, even during inflation and monetary restraint, because of a "liquidity leak of money substitutes" which occurs as monetary restraint is applied.<sup>7</sup> As  $L$  approaches infinity, for example, equation (21) approaches equation (1).<sup>8</sup> This Eisner view as to interest responsiveness of money demand is a minority one, however.

### REFERENCES

- W. H. Branson and A. K. Klevorick, "Money Illusion and the Aggregate Consumption Function," *Amer. Econ. Rev.*, Dec. 1969, 59, 832-49.
- R. Eisner, "Factors Affecting the Level of Interest Rates: Part II," in United States Savings and Loan League, *Savings and Residential Financing, 1968 Conference Proceedings*, Chicago 1968, 28-40.
- , "Fiscal and Monetary Policy Reconsidered," *Amer. Econ. Rev.*, Dec. 1969, 59, 897-905.
- G. Fromm and P. Taubman, *Policy Simulations With An Econometric Model*, Washington 1968.
- D. E. Laidler, *The Demand for Money*, Scranton 1969.
- D. Patinkin, *Money, Interest, and Prices*, 2d ed., New York 1965.

<sup>7</sup> See Eisner (1968) pp. 33-37, for a more complete statement of this view.

<sup>8</sup> For this case, make  $M_d = mY - L(r - u)$ , with  $M_d$  and  $Y$  as money variables, and  $u$  equal to  $r_1$ , so that the  $LM$  schedule is horizontal at  $r_1$ . Initial equilibrium  $Y$  is then as  $Y_b$  in Figure 2.

# Fiscal and Monetary Policy Reconsidered: Further Reply

By ROBERT EISNER\*

In the first part of my "Fiscal and Monetary Policy Reconsidered," I dealt in purely fiscal terms and argued, with particular application to the 1968 tax surcharge, that the effects of temporary changes in income tax rates would be limited to the extent that perceptions of permanent income were not altered. In the model presented by Jack Vernon, the effects are even more limited because 1) the initial round of decreased spending consequent upon the tax increase causes equi-proportionate declines in money income and prices but no reduction in real income and hence no further reductions in spending, real or nominal, and 2) with a given nominal money supply, the extent of the price decline is further diminished by the inflationary effect of a decline in the rate of interest.

It is hard to know what all this has to do with "money illusion." Vernon's nonhomogeneous expenditure function in money terms is his invention, not mine. Vernon's world of prices perfectly and instantly flexible downward from some point of inflationary equilibrium to a point of noninflationary full employment is also his rather than mine.

It is true that in the first section of my paper I made no distinction between changes in money and real variables. I assumed implicitly that spending units perceive cuts in money income as cuts in real income and have marginal propensities to consume which relate to those of real consumption functions. It is difficult to see that any other assumption is relevant unless we are to stipulate away both reality and the issue. The question is precisely whether or how much expenditures will be reduced as a consequence of an increase in tax rates. Depending on this reduction in expenditures will be a subsequent reduction in prices as sellers find goods unsold at existing levels. The extent to which

(relative) reductions in nominal effective demand will involve reductions in real output and the extent to which they will involve reductions in prices is a difficult and rather painful question both for theorists and the current Administration. The evidence of the last several years has been that the preponderant if not total effect has been on real output and not on prices.

What Vernon in effect argues is that *if* the deflationary tax policy had succeeded in lowering prices rather than real income, it would have lowered them with no multiplier effect; that is, by no more than the first round impact of reduced nominal spending. Vernon's reconsideration of anti-inflationary fiscal policy thus goes further than mine. While I would have room for a normal multiplier for a small initial multiplicand of tax-induced reductions in real consumption, Vernon would reduce the multiplier to one while retaining a small initial multiplicand which would spend itself entirely in a reduction in prices.

As to Vernon's second argument about the role of monetary factors, simply enough, I was in this section of the paper considering pure fiscal effects and operating with the implicit but I would have thought fairly standard and obvious assumption that monetary policy would be accommodating (although it can quickly be admitted that this is not always so). One would, of course, be correct in drawing from Vernon's equations the conclusion that a monetary authority which allows real cash balances to rise while the fiscal authority is trying to combat inflation by raising tax rates will tend to reduce the anti-inflationary impact of fiscal policy. In the second section of my paper, it will be recalled, I did consider the role of money and monetary policy, and while I remain inclined to downgrade its potency even more than that of fiscal policy, I do not argue that money generally does not matter at all.

\* Northwestern University and National Bureau of Economic Research.

# Pollution and Pricing

By ALLEN V. KNEESE\*

Jerome Stein's article in a recent issue of this *Review* is subject to misinterpretation which could lead to seriously misleading results. He makes several arguments against the Council's conclusion that "Many, though not all, pollution problems are local in character, and therefore determination of the appropriate level of environmental quality in these cases is likely to be more accurate if it is done locally rather than by the Federal Government . . ." (quoted by Stein, p. 532). One of his arguments is that local pricing (effluent charges) would result in inefficiency because the marginal product of pollution would then differ from region to region.

Possible confusion results from the fact that two separate measures of pollution are not consistently distinguished. The first is a physical measure like pounds of biochemical oxygen demand or tons of sulfur oxide discharge. The other is the dollar value of damage. In Stein's Figure 1, the horizontal axis represents pollution *damage* and what is demonstrated is that a dollar's worth of damage must be charged for at the rate of a dollar if efficiency is to prevail. This is emphatically not the same as saying that every pound of discharge should be charged for at the same rate, no matter where in the nation it occurs. But the latter is a conclusion that could easily be inferred from Stein's discussion. For instance, Stein says, "For example,

if the firms were ordered to reduce the rate of pollution by  $\Delta x$ , . . ." (p. 533) where the correct reference is to pollution *damage*. A similar sliding over from pollution emission to pollution damage occurs again in the first full paragraph of page 535.

The translation of pounds of discharge into dollars of damage depends, among other things, on meteorological and hydrological conditions and on the presence and preferences of receptors. Charging a dollar for a dollar's worth of damage requires quite different rates of charge per pound of discharge depending upon the locality. A uniform national charge might be defended on other grounds (see Kneese) but not as a requirement for efficiency strictly speaking.

Charges which reflect local conditions could in principle be set locally or nationally. If the Council's argument is right, and I am not trying to assess it here, that the optimum level of environmental quality is more likely to be accurately determined locally, then, problems of implementation aside, more efficient charges could be set locally.

## REFERENCES

- A. V. Kneese, "Environmental Pollution: Economics and Policy," *Amer. Econ. Rev. Proc.*, May 1971, 61, 153-66.  
J. L. Stein, "The 1971 Report of the President's Council of Economic Advisers: Micro-Economic Aspects of Public Policy," *Amer. Econ. Rev.*, Sept. 1971, 61, 531-37.

\* Resources for the Future, Inc. I am indebted to Anthony Fisher for discussion about this note.

# Local versus National Pollution Control: Note

By SAM PELTZMAN AND T. NICOLAUS TIDEMAN\*

In a recent issue of this *Review*, Jerome Stein criticized the discussion of public policy on pollution contained in the 1971 Report of the President's Council of Economic Advisers. His paper contains logical errors. It also obfuscates an important policy problem raised in the Report: What institutional arrangement can produce information about pollution costs and benefits most efficiently?

The burden of Stein's argument is that local autonomy in the setting of pollution charges will lead to differences in these charges, that the existence of different prices for the same service (use of the environment) violates the production and consumption conditions for a Pareto optimum, and that therefore the pricing decision (or, what is the same in this context, the decision as to how much pollution to tolerate in a given locale) should devolve upon the federal government. Stein's analysis ignores the essential spatial aspects of the pollution problem, fails to distinguish between the desirability of a single price and that of having a single price setter, and fails to distinguish between the relevant short and long runs. We contend that nationally uniform pollution charges could only be optimal in the long run in a very unusual world and particularly are not optimal in this world in the short run (now), and further that a temporally efficient set of charges is more likely to emerge under local rather than federal control. These contentions should be understood to apply to the context in which the Council's report recommends local autonomy, namely "where most benefits and costs of pollution are borne locally" (p. 121).

We deal first with the spatial aspects of the pollution problem, if only because Stein's analysis suppresses them. His economic sys-

tem of spatially uniform equilibrium prices for all goods and factors would require either that transport costs were zero, or that land were homogeneous with nonincreasing returns to scale in all production functions at all levels of output—i.e., no cities.<sup>1</sup> Where economies of scale and heterogeneous land create clusterings of economic activity (where cities are efficient), there will be spatially nonuniform demands for productive inputs including environmental inputs. As long as the cost of transporting these inputs is positive, there will be spatial differences in their prices in equilibrium.

In an economy with cities, spatially uniform prices for environmental resources will be particularly inefficient because population tends to cluster in the same areas as economic activity. Since pollution yields both benefits (in production) and costs (to residents), *both* the demand for use of environmental resources and the marginal social costs associated with their use will be higher at every level of pollution, implying a higher price the greater the density of economic activity. Indeed, Stein would have been more nearly correct had he argued for spatially uniform environmental standards (permissible pollution levels) rather than uniform pollution charges. If one abstracts from all regional variation other than population-production density, optimal pollution levels would be roughly the same everywhere, and the associated pollution charges would vary proportionately with density. This argument is illustrated in Figure 1, where schedules  $B_I$  and  $C_I$  represent, respectively, the marginal benefits to polluters and the marginal costs to residents of various levels of pollution in one location. If another place has double the density of population and pollution-generating activity, its marginal benefit schedule,  $B_{II}$ , will be to the *right* of  $B_I$  by a factor of two (the benefits of pollution being private),

\* University of California, Los Angeles and Harvard University, respectively, (order of the names chosen by a random process). We were Senior Staff Economists, Council of Economic Advisers, 1970-71. Our defense of the Council's Report should not, however, be construed as necessarily reflecting the views of the Council.

<sup>1</sup> See Edwin Mills, pp. 198-200.

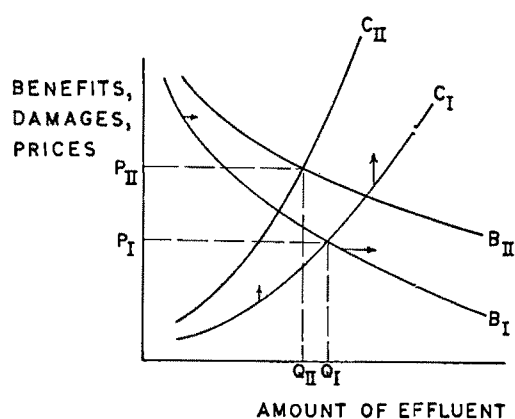


FIGURE 1

while the marginal cost function,  $C_{II}$ , will lie above  $C_I$  by a factor of two (the costs of pollution being public).<sup>2</sup> The net result will be approximately equal equilibrium quantities of pollution in the two places (exactly equal if the marginal benefit function is a rectangular hyperbola) and an equilibrium price,  $P_{II}$ , about twice as high as  $P_I$ .

To put the proposition in more intuitive terms: With its higher population density, New York City suffers more damage from a given level of pollution than Providence, Rhode Island. However if the price of pollution is raised accordingly, New York's air may still be about as dirty as Providence's, because of the greater density of pollution-generating activity in New York. It must be emphasized that this is an approximate result that abstracts from regional variations in climate, topography, income, etc. any of which might require regional variations in both prices and amounts of pollution.

If Stein's critique of the Council's Report fails to account properly for equilibrium differences in the marginal cost of pollution, it is also invalid even where such differences would disappear in equilibrium, so that Pareto optimality would require a spatially uniform pollution price. First, simply as a logical matter, a uniform price does not require federal responsibility for pollution charges. Indeed, reference to the simplest competitive

model should be sufficient to convince one that the tendency toward a uniform price is not inconsistent with decentralized ownership of the good being traded. Moreover, an important rationale for decentralized decision making is precisely that such a regime will enable the *optimum* price to be discovered and established at lowest cost.

This elementary principle has an important bearing on the policy problems with which the Council's Report was trying to deal. The political pressure for more pollution control is symptomatic of a growing awareness that prevailing property rules have resulted in too low a price on use of the environment. Before this disequilibrium can be eliminated, however, the equilibrium pattern of prices or standards has to be discovered. Given prevailing rules for use of the environment, which are frequently so ill-defined that no price is placed on use of the environment, information on the correct equilibrium price is not likely to come cheaply. The basis of the Council's recommendation of local control is that this will minimize the relevant information cost:

Since these rules require that the gains and losses entailed by different levels of environmental quality be weighed, the Government agency making the rules must be responsive to those who bear the gains and losses. This is especially important because part of the damage from pollution cannot be measured directly but depends on such things as the aesthetic preferences of those affected. As a practical matter, much of the damage from pollution will be 'measured' by political pressures from those damaged. [p. 121]

The Council's Report may be wrong or naive to suppose that evaluation of costs and benefits of pollution will be more accurate if this is made the responsibility of the localities experiencing the benefits and costs. Stein, however, gives us no reason for believing that federal control will lead to superior evaluations. Indeed, by jumping from the conclusion that there should be a single national price to the conclusion that there should be one national price setter, Stein ignores com-

<sup>2</sup> For a discussion of vertical summation for public goods, see Paul Samuelson.

pletely the problem of how this price setter is going to determine the correct price. Costless information is not characteristic of the real world nor was it assumed to be in the Council's Report.

To meet Stein on his own terms, however, even in a zero-information-cost world with a spatially uniform equilibrium price for pollution, a single national price would not be efficient in the transition from the current disequilibrium to the long-run equilibrium price. To elaborate, consider the following case which characterizes the situation of concern to the Council's Report and Stein's critique: Assume that in two localities (I and II) there is initially a zero pollution price.

Other factor prices and production conditions are identical, so that a single schedule of marginal benefits from emitting effluents into the environment is valid for both localities. This is illustrated in Figure 2 as the

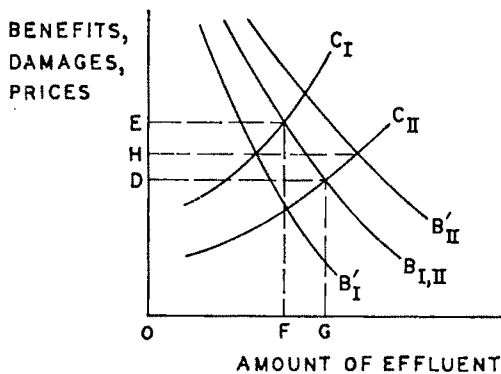


FIGURE 2

curve  $B_{I,II}$ . However, because of differences in climate, topography, etc., the marginal pollution damage functions differ, with that in I ( $C_I$ ) lying above that in II ( $C_{II}$ ). Stein correctly points out that, under local control, net-benefit-maximizing local governments would set different prices in this situation, specifically  $OE$  in I and  $OD$  in II.<sup>3</sup> However, so long as it is sufficiently costly for resources committed to effluent-generating ac-

tivity to relocate, a net-benefit-maximizing federal government would establish the same prices. The fact that there will then be a difference between I and II in the marginal rate of substitution between the environment and other goods is of no relevance until the cost of relocating pollution sources is sufficiently low to yield net gains from interregional "trade" in pollution. However, once this pair of short-run equilibrium prices ( $OE$ ,  $OD$ ) has been established long enough for relocation costs to be sufficiently small, the resulting net gains from trade will be exploited. In response to the higher price in I, effluent-generating activity will tend to relocate from I to II. As this relocation takes place, the marginal benefits function will shift rightward in II and leftward in I. These shifts will cause the net-benefit-maximizing pollution charges in I and II to converge.

If there were no long-run relocation costs, a long-run equilibrium would be attained with a uniform national charge ( $OH$ ) somewhere between  $OE$  and  $OD$  and with new short-run marginal benefit functions  $B'_I$  and  $B'_{II}$ . Stein is simply incorrect when he asserts: "Such a situation [shift of pollution] would not occur if the regulation of the use of the environment were the prime responsibility of the locality" (p. 535). So long as each locality is concerned solely with maximizing its own net benefits, pollution will shift from I to II, pollution charges will fall in I and rise in II until a long-run, Pareto-efficient equilibrium is attained. In making these convergent price adjustments, neither locality I nor locality II need be concerned with or know about the optimum price in the other locality. They need each only react to the resource shifts as they are experienced in each locality. Further, a national-net-benefit-maximizing federal government would not impose a different temporal price pattern, even if it knew the long-run equilibrium price at the outset. It would wish to encourage pollution to shift from I to II (at a cost minimizing rate) and, in the absence of some superior reallocation mechanism, this would entail high and falling prices in I, low and rising prices in II until long-run equilibrium is attained.

<sup>3</sup> Or they would auction rights to differing amounts of pollution,  $OF$  in I and  $OG$  in II.

Once it is clear that Pareto-efficient pricing of the environment is consistent a priori with either local or federal control, the relevant issue becomes, again, which form of control is likely to yield the more accurate evaluations of benefits and damages. The Council's report, as has been mentioned, relies on the presumed information-cost advantages of local self-interest in opting for local control. Stein adduces two reasons why enlightened provincialism may generate inaccurate information:

1. *Firms may Exercise Monopsony Power to Prevent Localities from Optimizing* This is, of course, possible, but Stein gives us no reason for supposing it to be probable. It is also possible, and perhaps even more likely, that governments may exercise monopoly power in the sale of pollution rights to the firm. Firms are only occasionally monopsony buyers of labor within any government jurisdiction. However, governments are always monopoly sellers of pollution rights within their jurisdictions. The importance of this monopoly power will, of course, depend on the cost of alternative locations open to the firm. The more centralized the control over sale of these rights, the fewer the alternatives and the greater the probability of inefficiently high prices for pollution rights. We do not see why competition in the sale of pollution rights has anything less to recommend it on logical grounds than, for example, competition in the sale of industrial sites or any other resource.

2. *The Environment is Often a Public Good.* Stein asserts that, since people all over the country benefit from the New England coastal view, the residents of New England should not be permitted to sell their view for their own higher income. One cannot, however, jump from the existence of some externality to the policy conclusion reached by Stein without some notion of the importance of the externality. The limitations of Stein's approach are best seen by analogy. Suppose that a majority of the nation feels that Indian tepee villages are a scenic treasure. Should we then prohibit Indians from living in houses? Most people would say no. Even though most people might prefer that some Indians

live in tepees, the stake that Indians have in their own housing is far greater. More generally, unless there is a perfect market in votes, a democratic process must have voting power distributed in the same manner as the consequences of a decision, to avoid bias.<sup>4</sup> It is important that the unit of local government making pollution-control decisions encompass nearly all parties significantly affected by pollution and that it not include a large proportion of persons insignificantly affected. In some areas there may be no level of local government that meets these criteria. But unless the relevant externality is significant nationwide, a national electorate would probably set much worse prices for pollution than any level of local government, because a majority of the national electorate would bear only a small fraction of the consequences. Where a local environment is a public good nationally, there may be a case for a tax on localities that permit pollution, or a subsidy for those that do not, but there is no general case for national controls. Since *some* externality inheres in almost any transaction, Stein's argument would leave almost nothing for local or individual control.

In summary, the existence of a nationally uniform pollution price is not a requirement for a Pareto optimum even in long-run equilibrium, and attainment of efficient prices is not inconsistent with local control over pollution prices. Indeed, so long as the effects of pollution and pollution control are borne locally and each locality maximizes its own net benefits, the temporal price pattern that emerges will be dynamically efficient. The real question is whether local or national control will produce the more accurate evaluation of pollution costs and benefits. Any general presumption about the information efficiency of decentralized resource control, we would submit, argues for local control. To opt for national control risks converting a possible local monopsony problem into a more probable national monopoly problem, and opting for national control because of some

<sup>4</sup> More precisely, the median discrepancy between marginal cost shares and marginal benefit shares must be zero. See Tideman.

concern for a community's environment on the part of people who do not reside in that community is, we submit, inefficient overkill.

## REFERENCES

- E. S. Mills, "An Aggregative Model of Resource Allocation in a Metropolitan Area," *Amer. Econ. Rev. Proc.*, May 1967, 57, 197-210.
- P. A. Samuelson, "A Diagrammatic Exposition of a Theory of Public Finance," *Rev. Econ. Statist.*, Nov. 1955, 37, 350-56.
- J. L. Stein, "The 1971 Report of the President's Council of Economic Advisers: Micro-Economic Aspects of Public Policy," *Amer. Econ. Rev.*, Sept. 1971, 61, 531-37.
- T. N. Tideman, "The Efficient Provision of Public Goods," in S. Mushkin, ed., *Public Prices for Public Products*, Washington 1972.
- U.S. Council of Economic Advisers, *Economic Report of the President*, Washington 1971.

# Behavior of the Firm Under Regulatory Constraint

By JEROME L. STEIN AND GEORGE H. BORTS\*

Harvey Averch and Leland Johnson developed a theory of the monopoly firm seeking to maximize profit, but subject to a rate of return constraint imposed by a regulatory authority. It is assumed that the regulated rate of return exceeds the market rate of interest but is below the unregulated rate of return. They concluded that: (a) the firm does not equate marginal rates of factor substitution to the ratio of factor costs and hence does not minimize the cost of producing the output it selects; and (b) the firm has an incentive to expand into other markets, even if it operates at a long-run loss in those markets. Most of the subsequent discussion of this original and stimulating paper has examined the following propositions.<sup>1</sup>

1. The scale of the firm, (i.e., its stock of capital) rises as the regulated rate of return declines towards the market rate of interest.
2. The regulated firm will produce a larger output than it did when it was not regulated.
3. The regulated firm will adopt input proportions different from those that minimize the cost of the final level of output.
4. The capital-labor ratio of the regulated firm will exceed the ratio that minimizes costs for the final level of output produced.
5. The capital-labor ratio of the regulated firm will be larger than that which prevailed when the firm was unregulated.

Each article has been illuminating and technically elegant. In this paper we relate the propositions concerning the effects of rate of return regulation to the traditional techniques of analysis based upon the Viner-Wong envelope. By using this powerful

method of analysis, we can answer every question raised to date in a simple manner.

## I. Unconstrained Long-Run Equilibrium

Figure 1 describes a monopoly which is in long-run equilibrium. The demand curve is  $D$ , marginal revenue is  $MR$ , long-run average cost excluding rent is  $LAC$ , and long-run marginal cost is  $LMC$ . The scale of plant is measured by the capital stock  $K$ . In long-run equilibrium when the scale is  $K_0$ , the short-run average cost excluding rent is  $AC(K_0)$  and the corresponding short-run marginal cost is  $M(K_0)$ . This unregulated firm will produce an output  $q_0$  and charge a price  $p_0$ . It is assumed in Figure 1 that, in the neighborhood of equilibrium, there are increasing returns to scale so that  $LAC$  is declining and  $LMC$  is below the long-run average cost curve. The analysis would be unchanged if we assumed that the  $LAC$  curve was constant or

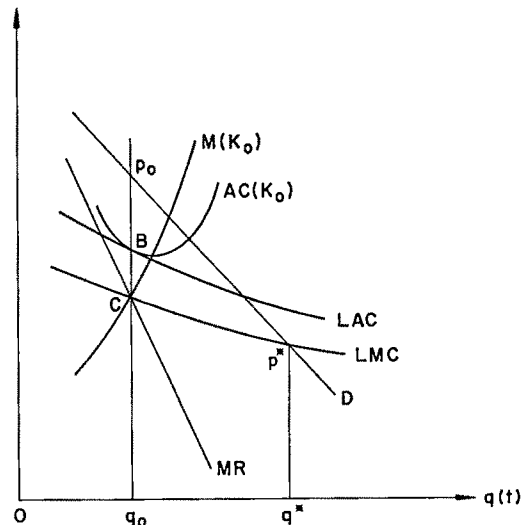


FIGURE 1. THE UNREGULATED MONOPOLIST IN LONG-RUN EQUILIBRIUM. OUTPUT  $q_0$  IS PRODUCED AT MINIMUM COST, SINCE SHORT-RUN AND LONG-RUN MARGINAL COSTS ARE EQUAL. PROFITS ARE MAXIMIZED SINCE THE LONG-RUN MARGINAL COST CURVE HAS AN ALGEBRAICALLY GREATER SLOPE THAN THE MARGINAL REVENUE CURVE AT POINT C.

\* Department of economics, Brown University. We are indebted to the following for excellent criticism of an earlier draft: H. Averch and L. Johnson, Elizabeth Bailey, W. Baumol, A. Klevorick, E. Sheshinski, A. Takayama, R. Weckstein, S. Wellisz, F. Westfield, R. Wichers and E. E. Zajac. The analysis of Westfield is similar to ours but appeared between the current and previous draft of this paper.

<sup>1</sup> See the bibliography at the end of this paper.

rising. In fact, it would be easier to analyze this problem if we assumed that there were constant returns to scale.

At output  $q_0$  total costs of production are minimized. Since the firm is on its envelope  $LAC$ , the marginal cost of producing output is the same whether a small increment of labor or capital is used to produce it. Both short- and long-run marginal costs are equal at point  $C$ . This means that the ratio of the marginal product of labor to the marginal product of capital is equal to the ratio of the marginal labor cost to the marginal capital cost. If (as we assume) the firm has no monopsony power, the latter is the ratio of the nominal wage to the cost of capital. Total profits are maximized since marginal cost is equal to marginal revenue; and long-run marginal cost exceeds marginal revenue for output greater than  $q_0$ . The firm earns a rate of return  $r_0$  which exceeds the market rate of interest  $i$ .

If the rest of the economy is competitive, then the opportunity cost  $C$  of the marginal resources used by this firm is less than the price of the product sold to consumers  $p_0$ . This means that the marginal rate of substitution in consumption between this good and other goods differs from the marginal rate of transformation in production. Output  $q_0$  does not correspond to a point on the economy's utility possibility frontier. If variations in  $q$  (to  $q^*$ ) exert no appreciable effect upon factor prices and incomes, then output  $q^*$  does correspond to a point on the utility possibility frontier.

## II. The Effects of a Regulatory Constraint

### A. General Considerations

Assume that this firm is subject to a regulatory constraint that it can earn no more than rate of return  $s$ , which exceeds the market rate of interest  $i$  but is less than  $r_0$ , the unregulated rate of return. Except for the constraint, the firm selects its rate of output to maximize profits. What happens to the input and output choices of the firm under those conditions?

Figure 2 is the same as Figure 1 without the long-run cost curves. Short-run average cost curve  $AC(q; K_0, i)$  is a function of out-

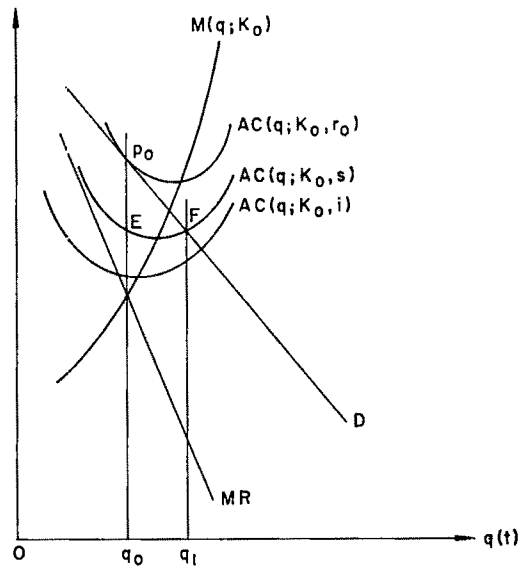


FIGURE 2. BY VARYING THE PRICE OF CAPITAL FROM  $i$  TO  $s$  TO  $r_0$  THE SHORT-RUN AVERAGE COST CURVE RISES ALONG THE SHORT-RUN MARGINAL COST CURVE  $M(q; K_0)$ : FROM  $AC(q_0; K_0, i)$  TO  $AC(q_0; K_0, s)$  TO  $AC(q_0; K_0, r_0)$ . THE STOCK OF CAPITAL IS  $K_0$ .

put and is evaluated at a cost of capital  $i$  and a scale of  $K_0$ . If price is equal to  $AC(q; K_0, i)$  then the firm earns a yield on capital equal to  $i$ . After paying its interest on bonds, it would have zero profits. If the capital were priced at  $s$ , then the average cost curve would slide upwards along the short-run marginal cost curve  $M(q; K_0)$  to  $AC(q; K_0, s)$ . Total variable costs would be unchanged, but total fixed costs would rise to  $sK_0$  from  $iK_0$  when capital is priced at  $s$  rather than at  $i$  and the scale of plant remains at  $K_0$ . If capital were priced at  $r_0$ , then the average cost curve would slide up the short-run marginal cost curve to  $AC(q; K_0, r_0)$ . At price  $p_0$  and output  $q_0$ , the firm would earn total profits of  $r_0K_0$  before payment of interest.

We assume that three conditions are met by the firm to satisfy the regulatory constraint in this context;

- 1) The firm must select a rate of output which will maximize its profits.
- 2) The firm should not earn a return on its capital in excess of  $s$ , where  $r_0 > s > i$ . Assume it earns  $s$ .

- 3) The firm must satisfy the quantity demanded at the price that is set.

The conditions cannot be satisfied when the scale of plant remains at  $K_0$ . If condition 2) is met, then price must be equal to average cost  $AC(q; K_0, s)$ . However, profits will only be maximized at price  $p_0$  which exceeds average cost. Therefore, conditions 1) and 2) cannot be met simultaneously with the scale of plant equal to  $K_0$  (the unregulated scale). If price were set at  $E$  and output were  $q_0$ , condition 3) would not be met. If price were set at  $F$  and output were set at  $q_1$ , then conditions 2) and 3) would be met, but condition 1) would be violated.

If the regulated rate of return were equal to  $r_0$ , then all three conditions would be satisfied. Profits are maximized at output  $q_0$ . Price  $p_0$  is equal to average cost  $AC(q_0; K_0, r_0)$  evaluated at a cost of capital equal to the regulated rate of return  $r_0$ . The quantity demanded at price  $p_0$  is  $q_0$ . This equilibrium occurs where the average cost curve, evaluated at a cost of capital equal to the regulated rate of return, is tangent to the demand curve at the price set. In this case where  $s=r_0$ , the regulation is ineffective and does not affect the economic behavior of the firm.

A moment of reflection and contemplation of Figure 2 will suggest what the equilibrium situation will be under an effective regulatory constraint. If a scale of plant  $K_1$  (greater than  $K_0$ ) were chosen so that the average cost curve  $AC(q; K_1, s)$ , evaluated at a cost of capital  $s$ , were tangent to the demand curve  $D$ , then conditions 1), 2), and 3) would be satisfied. Profits would be maximized, the firm would earn a return of  $s$  on its capital, and the quantity demanded would be satisfied at the price chosen  $p_1 = AC(q_1; K_1, s)$ . Graphically, the equilibrium under regulation will look like the Chamberlin monopolistic competition solution when the cost of capital is evaluated at the regulated rate of return. In the Chamberlin model, the long-run average cost is tangent to the firm's demand curve as a result of the entry of new firms. On the other hand, in the regulated firm model, the tangency of the short-run average cost to the demand curve is pro-

duced by variations in the scale of the firm. We now prove that this indeed will happen.

### *B. The Basic Propositions Proved*

#### *1. The Scale of Plant Will Increase*

The first question to be answered is: What will happen to the scale of plant as a result of regulation? What will be the relation between the postregulated scale  $K_1$  and the preregulated scale  $K_0$ ?

The unregulated firm in Figure 2, producing output  $q_0$  with plant scale  $K_0$ , is making a return  $r_0$  which exceeds  $s$  when price is set at  $p_0$ . The "excess" profits  $(r_0 - s)K_0 = (p_0 - E)q_0$  are not consistent with condition 2) of the regulatory constraint. We already showed that conditions 1) and 2) could not simultaneously be satisfied if the firm maintained scale  $K_0$ .

As long as the firm is earning a rate of return at least as great as  $s$ , total permissible profits  $sK$  will increase by increasing the stock of capital. This is a simple but crucial point. If unregulated profits  $rK$  exceed permissible profits  $sK$ , where  $r$  is the return to capital, then a rise in  $K$  will raise total profits  $sK$ . E. E. Zajac, and William Baumol and Alvan Klevorick showed that this is the heart of the problem.

The regulated firm will, therefore, increase its capital stock, as long as the average cost curve evaluated at a cost of capital  $s$ ,  $AC(q; K, s)$  lies somewhere below the demand curve. See Figure 2 for such a case. By expanding its scale, the firm will be able to increase its total profits  $sK$  which the regulatory agency will allow it to keep. The answer to the first question is that: The stock of capital will rise as a result of the imposition of a regulatory constraint.

#### *2. Assumptions Concerning the Viner-Wong Envelope*

The remaining propositions will only be correct if the traditional assumption is made concerning the Viner-Wong envelope: The short-run marginal cost curve shifts uniformly to the right as the scale of plant increases. This assumption is equivalent to the statement that: 1) a rise in the stock of capi-

tal raises the marginal physical product of labor. Moreover, we can infer the effect of regulation upon the capital intensity if we assume that: 2) the ratio of the marginal product of labor to the marginal product of capital is positively related to the capital-labor ratio. Any homogeneous production function will satisfy the second assumption. Figures 3a and 3b illustrate what happens to the short-run average and marginal cost curves when the scale of plant increases from  $K_0$  to  $K_1$ , given assumptions 1) and 2) above.

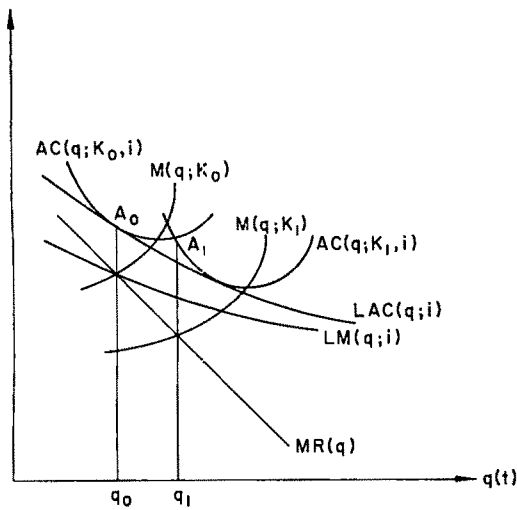


FIGURE 3a. WHEN THE SCALE OF FIRM INCREASES FROM  $K_0$  TO  $K_1$ , THE AVERAGE COST CURVE SHIFTS FROM  $AC(q; K_0)$  TO  $AC(q; K_1)$ . THE MARGINAL COST CURVE  $M(q; K_1)$  LIES UNIFORMLY BELOW THE MARGINAL COST CURVE  $M(q; K_0)$ .

Originally the firm possessed a plant scale  $K_0$  and produced an output of  $q_0$ . Production was efficient, since the firm was on its envelope. When it increases its scale of plant to  $K_1$ , the short-run average cost curve slides along the envelope to  $AC(q; K_1, i)$ . The average cost curves are evaluated at the market rate of interest  $i$ . The short-run marginal cost curve  $M(q; K_1)$  lies below  $M(q; K_0)$ , the short-run marginal cost curve associated with the short-run average cost curve  $AC(q; K_0, i)$ . Why?

At any rate of output (say  $q_1$ ), the short-run marginal cost is equal to the wage ( $W$ )

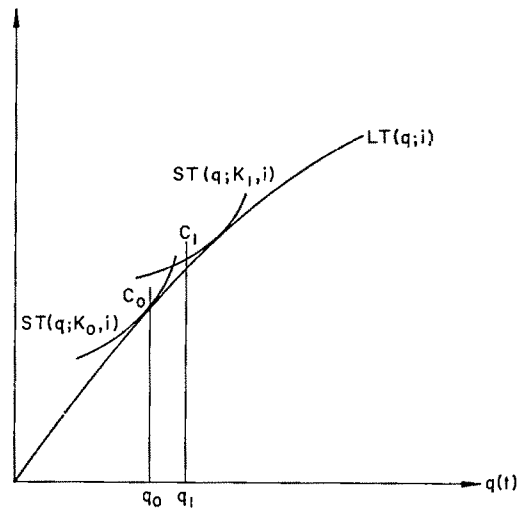


FIGURE 3b. THE LONG-RUN TOTAL COST CURVE IS  $LT(q, i)$ . THE SHORT-RUN TOTAL COST CURVE IS  $ST(q; K_0, i)$  WHEN THE SCALE OF PLANT IS  $K_0$ ; AND THE SHORT-RUN TOTAL COST CURVE IS  $ST(q; K_1, i)$  WHEN THE SCALE OF PLANT INCREASES TO  $K_1$ .

divided by the marginal product of labor ( $MPL$ ). If the stock of capital increases, then the *same output* can be produced by decreasing the employment of labor. The ratio of capital to labor will rise, and the marginal product of labor will increase. The short-run marginal cost ( $W/MPL$ ) of producing any rate of output  $q$  will decline as the scale of plant is increased from  $K_0$  to  $K_1$ . The short-run marginal cost curves  $M(q; K)$  cannot intersect when the assumptions 1) and 2) above are made. This is the situation described by the Viner-Wong envelope in Figure 3a. The remaining propositions are immediately deducible from this conventional analysis.

### 3. The Rate of Output and the Price

We have shown that the scale of plant will rise above  $K_0$ , as a result of the imposition of the regulatory constraint. The assumptions imply that the short-run marginal cost curve  $M(q; K_1)$  lies uniformly below  $M(q; K_0)$ , since  $K_1$  exceeds  $K_0$ . Profits will be maximized where marginal revenue is equal to the new marginal cost  $M(q_1; K_1) = MR(q_1)$ . The marginal revenue curve is negatively sloped. Consequently, the equality of mar-

ginal revenue and marginal cost can only occur by increasing the rate of output above  $q_0$ . Since the demand curve is negatively sloped, the price will decline as output expands. When the stock of capital expands to  $K_1$ , output will increase to  $q_1$  and price will decline to  $D(q_1)$ .

We have thereby derived the second proposition: The regulated firm will produce a larger output at a lower price than it did when it was not regulated.

#### 4. *The Inefficient Use of Inputs: The Scale of Plant is Excessive*

The inefficiency of regulation follows very easily from the properties of the envelope. At the unregulated equilibrium, profits were maximized. Assume that this equilibrium position is unique. This means that the marginal revenue curve  $MR(q)$  lies below the long-run marginal cost curve  $LM(q)$  for all rates of output greater than  $q_0$ . As a result of regulation, output rises above  $q_0$ . At the new equilibrium short-run marginal cost  $M(q_1; K_1)$  is equal to marginal revenue. Therefore, short-run marginal cost  $M(q_1; K_1)$  is below long-run marginal cost  $LM(q)$ . See Figure 3a which shows that marginal revenue  $MR(q)$  is below long-run marginal cost  $LM(q)$  for levels of output above  $q_0$ .

At output  $q_1$  the firm is not operating on its envelope, which means that short-run average cost  $AC(q_1; K_1, i)$  exceeds long-run average cost  $LAC(q_1, i)$ . Proposition 3 is deduced: The regulated firm will adopt input proportions different from those that minimize the cost of producing the level of output chosen.

Figure 3b describes this situation in terms of total cost curves. At the unregulated point, short-run total cost  $C_0 = ST(q_0; K_0, i)$  is equal to long-run total cost  $LT(q_0, i)$ . Or, short-run average cost  $A_0 = C_0/q_0$  is equal to long-run average cost. Output  $q_0$  is produced in the cheapest possible way. When the scale of firm rises to  $K_1$ , then the relevant short-run total cost curve is  $ST(q; K_1, i)$ . At point  $q_1$ , short-run total costs  $ST(q_1; K_1, i)$  exceed long-run total cost  $LT(q_1, i)$ . That is to say that short-run average costs  $C_1/q_1 = A_1$  exceed long-run average costs  $LT(q_1)/q_1 = LAC(q_1, i)$ . See Figures 3a and 3b.

The source of the inefficiency is that the scale of plant  $K_1$  is too large for  $q_1$ , the output produced. Short-run total costs are above long-run total costs at output  $q_1$  and scale  $K_1$ .

Suppose that we wished to decrease the quantity of output produced by one small unit. Short-run marginal cost is the savings in cost that results when the variable input (labor) is decreased. Long-run marginal cost is the savings that results when both inputs can be decreased. Since long-run marginal cost exceeds short-run marginal cost, the marginal cost of producing output is greater when the capital input is adjusted than it is when the labor input is adjusted. At point  $C_1$  in Figure 3b, the slope of the long-run curve is greater than the slope of the short-run curve. We could reduce the total cost of producing output  $q_1$  by decreasing the input of capital and by increasing the input of labor. The marginal cost of the capital input exceeds the marginal cost of the labor input. It is quite apparent (in Figure 3b) that total costs of production can be decreased by decreasing the scale of plant below  $K_1$ .

#### 5. *The Capital-Labor Ratio is Too High for Efficient Production, and Exceeds the Ratio that was Used in the Unregulated Case*

Propositions 4 and 5 can be deduced from Figure 3a or 3b and the assumptions made above. Long-run marginal cost  $LM(q_1, i)$  exceeds short-run marginal cost  $M(q_1; K_1)$ . Therefore, the marginal cost of expanding output by using capital exceeds the marginal cost of the expanding output by using labor. The latter is the wage  $W$  divided by the marginal product of labor  $MPL$ ; and the former is the cost of capital  $i$  divided by the marginal product of capital  $MPK$ . Since  $M(q_1; K_1) = W/MPL$  is less than  $LM(q_1)$ , it follows that:

$$(a) \quad \frac{i}{MPK} > \frac{W}{MPL}$$

or

$$(b) \quad \frac{MPL}{MPK} \equiv \frac{MPL}{MPK}(k_1) > \frac{W}{i}$$

We assumed that the ratio of the marginal

product of labor to the marginal product of capital is positively related to the capital-labor ratio  $k$ . This result is implied by any homogeneous production function. Therefore, Proposition 4 is derived: The capital-labor ratio  $k_1$  used by the regulated firm exceeds the ratio required for economic efficiency. Recall that we assume that the firm has no monopsonistic power so that  $W/i$  is a parameter.

At the unregulated point, production was efficient. Short-run and long-run marginal costs were equal (at point  $C$  in Figure 1). Therefore, the ratio of the marginal product of labor to the marginal product of capital was equal to factor price ratio  $W/i$ . When output  $q_0$  was produced, the unregulated capital intensity was  $k_0$ .

$$(c) \quad \frac{MPL}{MPK}(k_0) = \frac{W}{i}$$

Proposition 5 follows from (b), (c), and our assumption that  $MPL/MPK$  is positively related to  $k$ . The capital intensity  $k_1$  under regulation exceeds the capital intensity  $k_0$  in the unregulated case.

## II. The Effect of Tightening the Regulatory Constraint

The equilibrium of the regulated firm can be described by Figures 3a and 4. Receipts less total variable costs are maximal when the scale of the firm is  $K_1$  and output is  $q_1$ . This difference, equal to  $sK_1$ , is divided between interest to bondholders  $iK_1$  and monopoly profits  $(s-i)K_1$ .

If the ownership of the public utility were sold in a highly competitive securities market, the purchaser could not expect to earn the monopoly profits. Suppose that the sellers exchanged their ownership rights for bonds, which become the liabilities of the new owners. No change would occur in the scale of plant or rate of output, since quasi rents were already maximal. After the sale, the new owners of the firm earn no monopoly profits. Interest payments to the previous owners absorb the former monopoly profits  $(s-i)K_1$ . As far as the new owners are concerned, total receipts are equal to total costs. Total costs are total variable costs plus  $sK_1$ .

Or, price is equal to average cost evaluated at a cost of capital  $s$ . This is the situation described in Figure 4: Profits are maximized and are equal to zero. Total receipts are equal to total variable costs plus interest payments to all bondholders.<sup>2</sup>

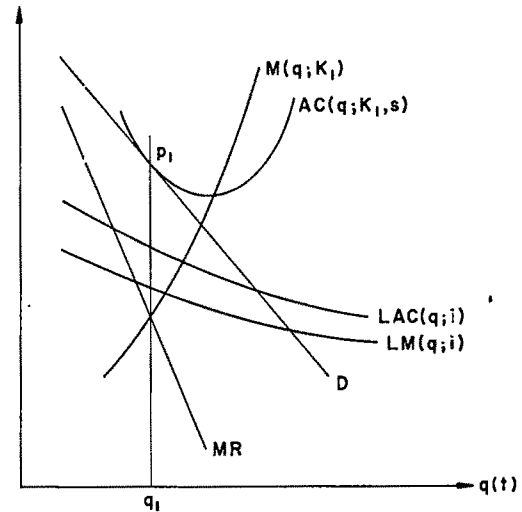


FIGURE 4. THE FINAL EQUILIBRIUM OF THE REGULATED FIRM. PRICE IS  $P_1$ , OUTPUT IS  $q_1$ . THE SCALE OF PLANT IS  $K_1$ .

The demand curve is tangent to the short-run average cost curve  $AC(q; K_1, s)$  evaluated at a cost of capital equal to  $s$ . Even if no outright sale of ownership occurred, Figure 4 would describe the postregulation equilibrium. Again, price would be equal to average cost evaluated at the regulated rate. Thus no matter what price is paid for the ownership of the firm, the appropriate long-run cost curves are  $LAC(q, i)$  and  $LM(q, i)$ .

<sup>2</sup> The conclusion that under certain circumstances a regulated firm will employ too high a capital-labor ratio and operate at too large a scale should not be confused with the conclusion that a particular firm does so. Anyone wishing to apply the Averch-Johnson analysis to the actual world of public utility regulation must find the answers to a number of empirical and theoretical questions specific to each case. For example,

- (a) Does the firm have monopsony power?
- (b) Do wages and capital costs reflect their opportunity costs in other industries?
- (c) What is the numerical value of  $i$ , the competitive cost of capital to this firm?

The equilibrium under regulation may, therefore, be found very easily. Start with Figure 2, the unregulated equilibrium where the firm makes  $r_0 K_0$  of monopoly profits.

(a) Calculate the short-run average cost curve  $AC(q; K_0, s)$  evaluated at the regulated rate of return. If monopoly profits  $r_0 K_0$  are capitalized as goodwill, the pre-regulated average cost curve would be  $AC(q; K_0, r_0)$ . Cost curve  $AC(q; K_0, s)$  would be a downward displacement of this curve along the marginal cost curve  $M(q; K_0)$ .

(b) Increase the scale of plant, thereby shifting the average cost curve  $AC(q; K, s)$  to the right, until it is tangent to the demand curve. Figure 4 is thereby obtained: The postregulated equilibrium.

What would happen if the regulatory constraint were tightened to  $s'$  where  $i < s' < s$ ? This question has already been answered, in effect. Going to Figure 2, the initial condition is now: output is  $q_1$ , scale is  $K_1$ , and rate of return is  $s$ . The firm is earning  $sK_1$  of quasi rents and can only keep  $s'K_1$ . Total profits can be increased by expanding the scale of firm above  $K_1$ . Hence, output and scale increase above  $(q_1, K_1)$ .

Graphically, we can see the effect of tightening the regulatory constraint by (a) sliding the average cost curve  $AC(q; K_1, s)$  in Figure 4 down the marginal cost curve  $M(q; K_1)$  to  $AC(q; K_1, s')$ , which is not drawn. Then: (b) increase the scale of plant, thereby shifting the average cost curve to the right. At the new and larger equilibrium scale  $K_2$ , (c) the average cost curve  $AC(q; K_2, s')$  will be tangent to the demand curve. Consequently, output will rise to  $q_2$  and price will decline to  $p_2$ . The conclusion is that a tightening of the regulatory constraint raises output and lowers price. Production remains inefficient insofar as an excessive amount of capital is used relative to labor, because the effective price of capital has been lowered by the regulation. Note that whatever the value of the regulated rate of return, (so long as  $s > i$ ) the appropriate long-run marginal cost curve of the firm is  $LM(q, i)$ .

We have shown how the Averch-Johnson conclusions can be explained on the basis of the traditional and powerful theory of the

firm which uses the Viner-Wong envelope. The reader, after going through this analysis, may feel that we have ignored two other cost curves of possible relevance to the problem: The envelope of the average cost curves when the return to capital is  $s$ , and the curve marginal to it. In terms of previous notation these would be labeled  $LAC(q, s)$ ,  $LM(q, s)$ . A review of Joan Robinson's profound chapter on four cost curves (pp. 133-42) will convince the reader that the curves inclusive of rent are irrelevant.

#### REFERENCES

- H. Averch and L. L. Johnson, "Behavior of the Firm Under Regulatory Constraint," *Amer. Econ. Rev.*, Dec. 1962, 52, 1053-69.
- W. J. Baumol and A. K. Klevorick, "Input Choices and Rate-of-Return Regulation: An Overview of the Discussion," *Bell J. Econ. Manage. Sci.*, Autumn 1970, 1, 162-190.
- E. Chamberlin, *The Theory of Monopolistic Competition*, 5th ed. Cambridge, Mass. 1947.
- I. Pressman and A. Carol, "Behavior of the Firm Under Regulatory Constraint: Note," *Amer. Econ. Rev.*, Mar. 1971, 61, 210-12.
- J. Robinson, *The Economics of Imperfect Competition*, London 1948.
- E. Sheshinski, "Welfare Aspects of a Regulatory Constraint," *Amer. Econ. Rev.*, Mar. 1971, 61, 175-78.
- A. Takayama, "Behavior of the Firm Under Regulatory Constraint," *Amer. Econ. Rev.*, June 1969, 59, 255-60.
- J. Viner, "Cost Curves and Supply Curves," in G. Stigler and K. Boulding, *Readings in Price Theory*, Chicago 1952, pp. 198-232.
- S. H. Wellisz, "Regulation of Natural Gas Pipeline Companies: An Economic Analysis," *J. Polit. Econ.*, Feb. 1963, 71, 30-43.
- F. M. Westfield, "Regulation and Conspiracy," *Amer. Econ. Rev.*, June 1965, 55, 424-43.
- , "Methodology of Evaluating Economic Regulation," *Amer. Econ. Rev. Proc.*, May 1971, 61, 211-17.
- E. E. Zajac, "A Geometric Treatment of the Averch-Johnson Behavior of the Firm Model," *Amer. Econ. Rev.*, Mar. 1970, 60, 117-25.

# The Futility of Pareto-Efficient Distributions

By E. J. MISHAN\*

Like Robin Hood and his merry men, the economist believes in taking from the rich to give to the poor. Talk of a "better" distribution of income, and he thinks at once of a more equal distribution in some sense. But apparently economists have never been quite happy simply to propose transfers of income to the poor as a commendable aim of policy. Regarding themselves as social scientists, it is apparently more satisfying to their search for status if the policy of equalizing transfer payments could be made to flow from some essential part of economics; if, that is, it could be given a sort of "scientific" underpinning.

The old fashioned way of going about this business is described in recent histories of economic doctrine, and we need only remind ourselves that with the help of some agreeable assumptions about the nature of the real world, the neoclassical Cambridge School was able to deduce equality of income as a corollary of maximizing society's total utility. By about 1939, however, scepticism about such propositions could no longer be contained. The old scaffolding was scrapped and that of the New Welfare Economics erected in its stead. Following some early years of heated debate, it became explicitly recognized that the new normative structure rested ultimately on a dual criterion, one involving allocation and one involving distribution. Admittedly, a dual criterion looks a bit untidy, and not surprisingly economists have been busy once more, this time trying to empty the contents of the distributive aspect into a purely allocative container—again making some agreeable assumptions about human nature.

## I. The Place of Distribution in the Old and New Welfare Economics

We need not spend much time in the old Cambridge School. A reading of Marshall and Pigou (1946), for instance, suggests that in fact they did believe that increased equal-

ity of incomes was in itself a good thing. In that event, recourse to assumptions of innately equal capacities for enjoyment and diminishing marginal utility of real income—necessary to ensure the conclusion that equality of real incomes maximized society's total utility—appears as an innocent attempt to give a scientific gloss to what was surely a strongly held value judgment. Had a utilometer been invented that proved beyond doubt that innate capacities for enjoyment varied in the most scandalous manner, being highest among the rich—in consequence of which maximizing total utility entailed an inequality of real incomes greater than that produced by the operation of the market—I have no doubt that their bluff would have been called. If the facts, that is, had disclosed a conflict between the "technocratic" goal of maximizing society's total utility and the more charitable goal of equalizing real incomes, it would have been the former that would have been thrown overboard.

Be that as it may, the advent of the New Welfare Economics in 1939, heralded by Kaldor and Hicks—though to some extent anticipated by Harold Hotelling and by Pigou himself—can be seen today as a movement aiming to dislodge the traditional concepts of cardinal utility (more particularly, diminishing marginal utility and interpersonal comparability) from the foundations of welfare economics, and to displace the criterion of utility maximization by a Pareto criterion. What is sometimes overlooked is the prolonged and intense, if narrow, concern it generated about the ethical premises of the new foundations and, indeed, of allocative statements in general. The significant consequence, however, was that after I. M. D. Little's *Critique* there could be no further pretence that judgments about economic

\* London School of Economics and the American University. I am grateful for the comments of a referee on a first draft of the paper.

efficiency were, in any acceptable sense, scientific. Efficiency statements were thenceforth to be understood as normative statements, and welfare propositions—whether allocative or distributive—were to rest ultimately on an ethical base.

Turning to more recent attempts to regard distributive implications as a form of externality, thus rendering distribution a part of allocative efficiency, we might like to remind ourselves that such effects—sometimes referred to as involving “interdependent utilities”—were given ample attention in numerous articles and also in several of the better known texts on welfare economics such as those of William Baumol, Little, and J. de V. Graaff (1957), and more recently by M. H. Dobb (1969) and S. K. Nath (1969). That none of these competent writers thought proper to invoke the interdependent utilities concept in order to dispose of the problem of distribution is not to be interpreted as an oversight on their part. Despite their unrelenting critiques of the subject and their frequent scepticism, they recognized that, for better or worse, welfare economics was an ethical study. Without ethical premises allocative statements, which are in effect prescriptive statements, have no significance for society. Once the ethical basis is acknowledged, however, one cannot legitimately determine the distribution of the product from a consideration of externalities. The reasons for this form the main thesis of this paper, and we turn to them in Sections III and IV. Section II, which follows, restricts itself to the auxiliary task of indicating that in any case (whether or not the reader is persuaded by the arguments in Sections III and IV) the attempts to derive distributional propositions from efficiency considerations is foredoomed to failure.

## II. The Impossibility of Efficiency-Derived Distributions

Although some recent writers, such as Richard Zeckhauser, have, in passing, treated distribution as a form of externality, their analysis and conclusions can easily be made independent of this assumption. We are con-

cerned more directly with papers devoted largely to arguing the case for translating directly from allocative efficiency to distribution. The most recent and ambitious attempt to do this has been made by Harold Hochman and James Rodgers (henceforth H-R), and since they exercise great care in presenting their case, I shall address myself primarily to their arguments.

In order to avoid any misunderstanding at the outset, let us be quite explicit that in talking of economic efficiency we elect to operate in the domain of normative economics. We are not, that is, restricting our analysis simply to explanatory hypotheses of political behavior as exemplified by the contributions of James Buchanan and Gordon Tullock, Anthony Downs, or Albert Breton, to mention only those that immediately come to mind. The very title of H-R's paper, “*Pareto Optimal Distributions*” [my italics] is clear indication of their concern with normative economics,<sup>1</sup> which concern is borne out in the text by such statements as “Efficiency criteria can be applied, therefore, to redistribution of income through the fiscal process,” or “Redistribution through the fiscal process is just as necessary for the attainment of Pareto optimality in these circumstances as the collective provision of conventional public goods” (p. 543).

In fact it would be rather unkind to interpret their paper otherwise—as an attempt at an explanatory hypothesis. For centuries, for millennia, the poor have received gifts in money and kind from those in more comfortable circumstances. The phenomenon goes by the name of charity and has long been conceived as a moral obligation which, when generously discharged, partakes of virtue. The economist having knowledge of such transfers might well regard it as a terminological advance to describe them as arising from “externalities,” specifically of the “interdependent utility” sort. But he surely does not deceive himself to the extent

<sup>1</sup> The same remark applies to Zeckhauser's paper, “*Optimal Mechanisms for Income Transfers*” (my italics).

of claiming that he is explaining the hitherto inexplicable, or that he is constructing a refutable hypothesis.<sup>2</sup> It is only fair then to continue to assume that H-R mean what they say, and that they are indeed intent on deriving distributional propositions from allocative criteria.

H-R's approach appears to be consonant with that of Kenneth Arrow insofar as any resultant ranking for society is to emerge from the specific orderings of the individual members. In Arrow's concept of the ranking of social states by each person, no exception is taken to the presumption that each person is likely to favor himself, his family, and his friends. H-R, however, narrow the problem to that of distribution and, by choosing appropriate assumptions about the utility interdependence of the rich Mutt and the poor Jeff, are apparently able to justify some transfer of income from Mutt to Jeff on the Pareto criterion. Apart from some minor ambiguity detected by Paul Meyer and J. J. Shipley, the mechanics of their demonstration are simple enough, and any further summary of their position serves no purpose here. Suffice it to say that theirs is an attempt

to persuade us that distribution need not be conceived as a separate aspect of a welfare criterion. Building on the benevolence of the rich, Pareto efficiency can be extended to the determination of distribution also. To quote them: "Both allocation and redistribution can be dealt with in terms of the same methodology and the same criterion—efficiency" (p. 543).

There are two main difficulties in any proposal to deal with distribution as a part of allocative efficiency:<sup>3</sup>

1. The operational value is practically nonexistent. In a community consisting of some scores of millions of adults, knowledge of the requisite pattern of interdependence is for all practical purposes unattainable. There can, perhaps, be some recourse to sample surveys. But to proceed from such surveys to distributional policies cannot be counted on to produce mutual benefits among all the members of the community. For a Pareto justification of redistribution policies cannot take the form that "on the average" or "on a balance of probabilities," a man earning \$20,000 a year would benefit from having, say, \$2,000 of it transferred to a man, any man, earning \$5,000 or less. The Pareto criterion, invoked by H-R, would require that each donor *individually* benefits from the transfer effected on his behalf. If, among the millions of citizens whose money is to be transferred to others (on the strength of an empirical investigation that comes up with

<sup>2</sup> Buchanan, in an impetuous moment perhaps, does seem to have claimed the latter. In talking of a "positive theory of political economy," he asserts that "a consistent methodological position does not allow the introduction of non-individualistic norms in *either* allocation or distribution." (presumably he meant a consistent *non-normative* methodological position). He goes on to argue, contrary to H-R incidentally, that "The mere fact that some members of the community are poor does not, in and of itself, normally impose an external diseconomy on many of the remaining members. What does impose such an external diseconomy is the *way* that certain persons behave when they are poor. It is not the low income of the family down the street that bothers us; it is the fact that the family lives in a dilapidated house and dresses its children in rags that imposes on our sensibilities. And we are willing to pay something to remove this external effect; it is relevant for behavior" (p. 189). Following this cynical interpretation of human motivation, he ends his paper with the resounding claim: "I advance the refutable hypothesis that distribution in kind is the predictable outcome of the political process" (p. 190). Since every one already knows that a good deal of distribution in kind has been and continues to be popular, this hypothesis is about as useful as the equally refutable one that people wear more clothes in winter.

<sup>3</sup> H-R, however, invite a gratuitous difficulty by acknowledging Richard Musgrave's point about a redistribution of income resulting from "the political power of the recipients" being something distinct from that arising from a Pareto efficiency criterion. Income transfers resulting from political power can be treated simply as a fact of political life or else they can be rationalized by reference to interpersonal comparisons of utility. However treated, it is not congruent with a Pareto efficiency criterion and could, indeed, directly conflict with it. This is not to deny that, following some politically motivated income transfers, there might remain scope for private benevolence and, therefore, for income transfers that are Pareto improvements. But if benevolence is to supplement political power, what weights are to be used in harmonizing them, and on what criteria?

average figures), any one protests simply that he is displeased with the prospect then there is no sanction for its transfer on the criterion proposed by H-R.<sup>4</sup> Nothing less than detailed information about each person will ensure the consent of everyone involved.

2. The validity of the concept is deficient. Even if all information were free and all transfers could be implemented without cost, Pareto-efficient redistributions cannot solve the distribution problem. This is so because any movement toward an optimum takes off from a given set of quantities and prices, and also a given distribution of assets (material and human). Such a movement toward an optimum therefore presupposes an initial distribution of real income. Thus, beginning with any particular initial pretax or gross distribution of real income in the community, there corresponds a number of Pareto optimal allocations and—taking into account any externalities of benevolence—a number of apparent Pareto optimal distributions also. However, there can be virtually an unlimited number of initial pretax distributions of real income. In consequence there can be virtually an unlimited number of apparent Pareto optimal distributions also. The problem of choosing among these resultant Pareto optimal distributions therefore remains. It cannot be solved without prior choice of a pretax income distribution. But this latter choice cannot be made on any efficiency criterion; only on a distributional one. Externalities of benevolence do not suffice.<sup>5</sup>

<sup>4</sup> True, H-R add that one might feel "... that the amount of redistribution dictated by the Pareto criterion will not be 'enough.' We are not saying that society should necessarily follow only the Pareto rule. It is possible, however, to develop a theory of redistribution based on such a rule ... " (p. 556). But the difficulty indicated in the text is not to be evaded by reference to the possibility of *other* distributional rules. As indicated in the preceding footnote, if some other rule is accepted, and it implies a different redistribution, is the Pareto optimal distribution to become inoperative? Are the two (or more?) distributions to be reconciled by a system of weights?

<sup>5</sup> It is perhaps unnecessary to remark that the authors' tables of hypothetical and actual distributions are evidence neither for nor against their optimal redistributions thesis. Existing tax structures can be "explained" in terms of political prudence, of political

Moreover, once we allow that efficiency considerations cannot solve the distribution problem for society, neither can we fall back on the more modest proposal of promoting Pareto-efficient redistributions regarded simply as Pareto improvements that start from a given distribution of real incomes. For once a "better distribution" has to be referred ultimately to some *non-Pareto* criterion, so also has any apparent "Pareto optimal" redistribution. Formally, then, we are faced once more with the problem of conflict between efficiency and distribution—here specifically with "efficiency-determined" distributions and "distribution-determined" distributions—which has exercised economists since 1939 and led them to a close examination of the logic of adopting a dual criterion.

### III. The Undesirability of Efficiency-Derived Distributions

Even if it were possible to propound an efficiency-determined distributional criterion, based on externalities arising out of income differentials, that was free of the above objections, it would run into ethical difficulties for at least two reasons.

First, pursuit of allocative efficiency—moving factors about until the value of their social marginal products become equalized—need entail no more than a *potential* Pareto improvement, not an actual one. As in a benefit cost analysis the criterion is met if gains exceed losses, which is to say if gainers *can* compensate losers. Once we introduce externalities into the allocative problem, then, we are bound to recognize the existence of "malevolence" as well as "benevolence." (If malevolence is too strong a word, perhaps "envy" is more acceptable as the response entailed by negative interdependent utilities.) Indeed, negative interdependent utili-

---

power, of interpersonal comparisons, of uncoordinated pressures within and without the civil service, of both pure and conditional benevolence, of some moral consensus, or of any combination of these, plus reservations on the quality of the statistics and the methods of imputation. H-R's explanation of the figures of the existing fiscal incidence runs in terms of benevolence, political power, and doubts about the data.

ties appear to have been more popular with economists than positive ones. Dusenberry's relative income hypothesis, for instance, has it that a man's utility rises the lower the income of others in relation to his own income. In a Mutt and Jeff world, the hypothesis would imply that, with a given real income, Mutt's utility would rise as Jeff became poorer. And even if the relationship were symmetric as between Mutt and Jeff, it would be decidedly awkward for the H-R thesis.

Suppose the facts about envy and benevolence in the real world are such that the pursuit of Pareto efficiency requires on balance transfers from the poor to the rich, what then? It is not to the purpose to appeal to the empirical evidence here even if it were conclusive; and it is not. If the facts were as hypothesized above, those adhering to efficiency distributions might be obliged to prescribe an income distribution more unequal than that emerging from the free play of the market. If in such a situation they demurred; if they restated their position more cautiously as one that lent support to efficiency distributions *provided* only that they would result in a less unequal distribution, then they would tacitly be acknowledging some other principle, presumably one bearing on equity.

Secondly, ignoring entirely the existence of nonbenevolent utility interdependences, the degree of benevolence within any income bracket is not, in general, uniform. This is so whether a donor's increase of welfare from a transfer of his income to those less fortunate than he is entirely independent of the donations of others, or whether the externality relationship is more complex—the donor's welfare being a positive function also of the amounts given by others (since the more given by others the larger the sums received by the poor).<sup>6</sup> But, allowing that all problems of information and implementation can be overcome, the uneven spread of benevolence can result in a Pareto optimal redistribution that would require some of the very rich to contribute next to nothing to the poor

and some of the middle income groups to contribute handsomely. A tax structure calculated to reflect such a pattern of benevolence, one that provides exemption for the nonbenevolent rich, is ethically inadmissible and is likely to be politically unfeasible.

#### IV. Concluding Remarks

The welfare economist has never sought to exclude considerations of distribution (which involve the welfare of others) from the individual's utility function. A member of society is seldom indifferent to the distribution of incomes at large, from which it follows that some kinds of distribution afford him more satisfaction than others. The question at issue then is not whether income distribution properly enters into a person's utility function. In general, it does so. The question at issue is how the economist is to deal with it.

If there are, as I have argued in the preceding sections, some formidable obstacles in the way of any attempt to determine the pattern of fiscal transfers wholly or partly by recourse to the externalities of benevolence, we might agree *faute de mieux* to continue with the traditional separation of allocation and distribution. The economist engaged, say, in a benefit-cost calculation would confine himself to standard allocative practice, and for the rest would at least point up any significant distributional implications. But there is, indeed, a stronger case for rejecting the interdependent utility approach to distribution—at least, if we continue to regard welfare economics as founded on ethics, not on utility. For there is much that might increase total utility, or that might realize Pareto improvements, that is nonetheless quite unacceptable to civilized societies and can, therefore, become no part of their agenda.

However much the aggregate utility enjoyed by a hysterical mob in kicking a man into insensibility exceeds the disutility of the victim, society would feel justified in intervening. Society would presumably also disapprove of a "mutually beneficial" agreement between a shopkeeper and a gang of toughs whereby the latter discontinues its practice of breaking the shopkeeper's win-

<sup>6</sup> The argument is developed by Robert Goldfarb.

dows in exchange for regular tribute. Again, though an anonymous donation of \$1,000 to some poor family so excited the envy of its neighbors as to entail a potential Pareto loss, society would be disinclined to withhold the donation. There is, alas, a great deal of envy about which, though it acts to reduce the welfare of those who learn of the good fortune of others, society need not and generally does not take into account in framing economic policy.

What I am asserting is that for his propositions and recommendations to have any relevance, the welfare economist must draw on the ethics of the society for which they are intended. Once it is conceded that welfare economics is founded on ethics, not utility, simplifications follow as a matter of course. For just as externalities arising from envy or malice may not be agenda for society, so neither may be externalities arising from benevolence. Benevolence, even if pure and evenly dispersed throughout society, would entail one set of transfers beginning from any pre-tax situation. The sense of justice would entail another. In a hypothetical state, that is, in which no man knows what his income in life is to be, or what position in society he will come to attain, there can be a disinterested debate on the limits within which incomes should range, and agreement reached on a tolerable *structure* of disposable incomes and, therefore, of the pattern of transfers necessary to implement it. And when men talk, neither of charity or benevolence, but of *distributional justice*, the authority they command varies with their success in projecting themselves into this hypothetical state that is prior to their existing worldly interests.

True, all members of society may not, at any given time, be wholly in agreement about the exact limits to be set on income variations, or about the shape of an ideal tax structure. For this reason the welfare economist, following public opinion in this respect, commits himself only to statements about the resulting distribution being 'on the whole' progressive or regressive. But it is

the sense of justice that impels society to reconsider the question from time to time and work toward some ideal.

#### REFERENCES

- K. J. Arrow, *Social Choice and Individual Values*, New York 1951.
- W. J. Baumol, *Welfare Economics & The Theory of The State*, London 1952.
- A. Breton, "A Theory of the Demand for Public Goods," *Can. J. Econ.*, Nov. 1966, 32, 455-67.
- J. M. Buchanan, "What Kind of Distribution Do We Want," *Economica*, May 1968, 34, 185-90.
- and G. Tullock, *The Calculus of Consent*, Ann Arbor 1962.
- M. H. Dobb, *Welfare Economics and the Economics of Socialism*, Cambridge 1969.
- A. Downs, "An Economic Theory of Democracy," *J. Polit. Econ.*, Apr. 1957, 65, 135-50.
- R. S. Goldfarb, "Pareto Optimal Redistribution: Comment," *Amer. Econ. Rev.*, Dec. 1970, 60, 994-96.
- J. de V. Graaff, *Theoretical Welfare Economics*, Cambridge 1957.
- H. M. Hochman and J. D. Rodgers, "Pareto Optimal Redistribution," *Amer. Econ. Rev.*, Sept. 1969, 59, 542-57.
- H. Hotelling, "The General Welfare in Relation to Problems of Taxation and of Railway and Utility Rates," *Econometrica*, June 1938, 6, 242-69.
- I. M. D. Little, *A Critique of Welfare Economics*, 1st ed., Oxford 1950.
- P. A. Meyer and J. J. Shipley, "Pareto Optimal Redistribution: Comment," *Amer. Econ. Rev.*, Dec. 1970, 60, 988-90.
- R. A. Musgrave, "Pareto Optimal Redistribution: Comment," *Amer. Econ. Rev.*, Dec. 1970, 60, 991-93.
- S. K. Nath, *A Reappraisal of Welfare Economics*, London 1969.
- A. C. Pigou, *The Economics of Welfare*, London 1946.
- R. J. Zeckhauser, "Optimal Mechanisms for Income Transfers," *Amer. Econ. Rev.*, June, 1971, 61, 324-34.

# Peasants, Procreation, and Pensions: Comment

By WARREN C. ROBINSON\*

In a recent article in this *Review*, Philip Neher presented a model aimed at showing how high fertility can occur because of the use by parents of children as investments to insure future consumption needs. This "pension effect" is not exactly a novel idea, but Neher has performed a modest service in providing a rigorous proof of this familiar proposition couched in the symbols of modern growth theory. But he also suggests that the exercise is meaningful for the real world and has policy implications. In fact, it can be shown that Neher's model is so partial a treatment and so naive with respect to likely empirical parameters, that it has nothing whatsoever to contribute to policy.

Taking plausible, real world data for the relevant variables, what is the probable strength of this pension effect? Fortunately, Goran Ohlin has already performed this exercise, and we can simply summarize his results. Ohlin's analysis is in terms of "adult-equivalent consumption units"; the amount of income required to support, at present standards, an adult male aged 25-30 for one year. It is based on empirical estimates drawn from studies of less developed countries (*LDCs*). He finds that a child living from birth to age 15 requires 8.9 adult-equivalent consumption units of subsistence, whereas an individual living from 60 to 85 requires 13.4 adult-equivalent consumption units of subsistence. The judgment of whether children can be economically used as a pension plan reduces them to a calculation of the tradeoff between the cost of supporting a child during the 0-15 age span compared to the parents' requirement during the 60-85 age span. Table 1 presents the results of this comparison. Ohlin argues that:

... a look at [the] table ... seems to lead to the conclusion that investment in children is a costly way of securing old-age support. The average cost of a

single child is on almost all assumptions greater than the needs of a single parent in old age. When the number of children is multiplied in order to increase the likelihood of having surviving children at hand, the disparity is enormously magnified. [p. 1728]

He concludes quite reasonably, that only under highly unlikely circumstances would children constitute a likely investment for the future. So much for the strength of the pension effect.

TABLE 1—PRESENT VALUES AT BIRTH OF CHILDHOOD CONSUMPTION AND OF OLD-AGE CONSUMPTION OF A PARENT AGED 25 AT BIRTH OF CHILD

	Consumption Units	
	Childhood (0-15)	Old Age (60-85)
A. Consumption of individual surviving respective age periods (undiscounted)	8.9	13.4
B. Same, discounted at		
1%	8.2	6.6
5%	5.9	0.4
10%	4.0	0.01
C. Average consumption per birth and per parent after depletion through mortality (undiscounted)	6.5	2.3
D. Same, discounted at		
1%	6.3	0.7
5%	4.7	0.05
10%	3.2	0.02

One or two other comments on Neher's assumptions are also in order. First, he argues that devising new, more inexpensive, or more acceptable birth control devices is essentially irrelevant to eliminating excess fertility, since people must *want* to reduce fertility, and if they do, then their existing methods, even very primitive ones, will do. This is a superficially attractive argument which is,

\* Pennsylvania State University.

however, wrong on two counts. First, the crux of the problem of excess fertility for many backward societies is that the traditional means of limiting population growth—infanticide, abortion, social controls on age at marriage, religious taboos requiring sexual abstinence on many holy days, bans on widows remarrying—have broken down. Fertility has typically been rising while mortality has been falling. Mortality itself was an effective check to population growth in premodern periods, and reduced death rates really go back less than a generation. Thus, the societies in question are in a state of transition—they have lost their traditional population control mechanisms and have not yet mastered the new, more purely chemical and mechanical ones developed in the West. Absence of information and of meaningful access to the modern devices explains failure to use them far more than does a lack of desire (see W. P. Mauldin).

In any case, it is strange for an economist to argue that making a service available at a lower price will not affect demand. The potential family planners can be thought of as having an implicit preference pattern for children versus other goods. Even if the net benefit of children is zero, they may still continue to have children, if it can also be shown that there exists a *cost* to preventing the births. The disutility of the children may be small, and the cost of preventing births (in time, money, and psychic terms) nonnegligible. Remember we are not talking only about the price of a package of condoms. We are talking about the psychic cost of a female submitting for the first time in her life to a gynecological examination; or the psychic cost of risking one's immortal soul. Finding ways of lowering or eliminating these costs is crucial.

Secondly, Neher argues that "excess" fertility is a special case of *market failure*, with

externalities caused by high fertility leading to the breakdown. But surely this is inconsistent with one of the important assumptions of his model. For if the family is viewed, as he suggests, as a self-contained, self-supporting economic unit with fixed resource endowment, then any costs of excessive fertility must be borne exclusively by that family unit. There will be *generational* effects but not *external* effects in the usual sense.

In fact, the true heart of the population problem does rest in the present or short-run future externalities imposed on society by the fertility decisions of some families. In the growing "population bomb" literature, no distinction is made between the externalities and the generational arguments, but they are different. The time-wise generation effects must remain uncertain and heavily dependent on the discount rate used. Present externalities represent the main, if not the exclusive, economic rationale for any population policy. (See B. Berelson.)

#### REFERENCES

- B. Berelson, "Beyond Family Planning," in *Stud. Family Planning*, Feb. 1969, No. 38, 1-39.
- W. P. Mauldin, "Fertility Studies: Knowledge, Attitude, and Practice," *Stud. Family Planning*, June 1965, No. 7, 1-15.
- P. A. Neher, "Peasants, Procreation, and Pensions," *Amer. Econ. Rev.*, June 1971, 61, 380-89.
- D. Nortman, *Population and Family Planning Programs: A Factbook*, Number 2, Reports on Population/Family Planning, Population Council, New York, June 1971.
- G. Ohlin, "Population Pressure and Alternative Investments," *International Population Conference Proc., London 1969*, Vol. III, International Union for the Scientific Study of Population, 1703-28.

# Peasants, Procreation, and Pensions: Reply

By PHILIP A. NEHER\*

Warren Robinson dismisses the strength of the "pension effect" on the grounds of Gorin Ohlin's data which suggest that children are a "costly way of securing old-age support." Quite apart from Robinson's confusion of marginal and average costs and returns, suppose the rate of return on children were so low as to be even negative. Should we then conclude that the pension effect has no force? Surely it would be naive to do so. Rates of return cannot be appraised *in vacuo*. For example, suppose nature offered people negative rates of return on children and everything else, and that intergenerational transfers are ruled out. Then parents, planning for their retirement, must invest at that negative rate, which will then be the observed rate, quite apart from their time preference. Only if children turn out to be bad investments, relative to other assets, can we conclude that the pension effect is weak. Even then, we might observe relatively low rates of return on children as assets, if they are thought to be relatively safe investments in diversified portfolios.

But if Robinson's conclusions still hold, if children have no place in a rationally selected retirement portfolio, then I am grateful to him for having highlighted the argument in Section IV of my article when I suggested

that a dominating "... good asset (bonds) drives out the bad asset (children)" (p. 387) so that "The population will wither away unless there are other motives for having children (p. 388).

Robinson goes on to question my fundamental assumption that human motivation has something to do with human fertility. Nothing he says convinces me that the desire to reduce family size is not an essential ingredient in the problem. The costs of fertility control can be viewed as costs associated with achieving a desired portfolio. If condoms, intrauterine loops, or other devices reduce these costs then I am in favor of them as much as is Robinson.

Finally I do not think Robinson and I really disagree about the nature of the "fundamental market failure" in my model economy. It was precisely intergeneration effects, within the family, I had in mind. But surely we do future generations a disservice if we ignore "time-wise generation effects."

If the postwar baby boom turns out to be a bulge on the population pyramid, then Robinson is shortsighted in discounting the relative magnitude of the intergeneration transfers when that bulge of babies becomes a bulge of dependent grandparents. Perhaps Robinson belongs to an unconcerned generation. After all the crunch will not come until 2010 A.D. or so.

\* University of British Columbia.

# A New Look at the Muth Model

By WYATT MANKIN\*

In his book, *Cities and Housing*, Richard Muth formulates a model of household locational choice. He assumes that the household works in a central business district, and it resides in one of the series of residential contours that surround the central business district. In the Muth model, the price of housing space is an inverse function of distance from the central business district, and the cost of commuting varies directly with both distance from the central business district and wage income. Muth finds that a change in wage income will increase commuting distance if the income elasticity of demand for housing space exceeds unity, but a change in non-wage income will increase commuting distance if the income elasticity of demand for housing space is positive.

A major limitation of the Muth model is that it lumps both leisure and goods into a single composite commodity. The purpose of this note is to demonstrate that new implications about the effect of a change in wage income on commuting distance can be obtained by distinguishing between leisure time and goods.

To show this let the household seek to maximize the following utility function:

$$(1) \quad U = U[X, H, L]$$

subject to the budget constraint:

$$(2) \quad 24w + y = P_X X + [P_h - P_h(k)]H + wL + T + T(k, w)$$

where:

- $X$  = a composite good
- $H$  = housing space
- $L$  = hours of leisure
- $w$  = hourly wage rate
- $y$  = nonwage income
- $P_X$  = price of composite good
- $P_h - P_h(k)$  = price of housing space
- $k$  = distance from the central business district
- $T + T(k, w)$  = expenditure on commuting

This specification of the household utility function and budget constraint differs from Muth's in two respects. First, the composite commodity of Muth's model is here dichotomized into both leisure and goods. Second in order to stress Muth's crucial assumption, (i.e., only changes in wage income alter the marginal cost of commuting) income is broken down into its' wage and nonwage components, and expenditure on commuting is made a function of wage income.

To find the household equilibrium conditions, set the first partial derivatives of the Lagrangian function:

$$(3) \quad M = U(X, H, L) + \lambda \{ 24w + y - [P_h - P_h(k)]H - wL - T - T(k, w) - P_X X \}$$

equal to zero. These derivatives are:

$$(4) \quad \partial M / \partial X = U_X - \lambda P_X = 0$$

$$(5) \quad \partial M / \partial H = U_H - \lambda [P_h - P_h(k)] = 0$$

$$(6) \quad \partial M / \partial L = U_L - \lambda w = 0$$

$$(7) \quad \partial M / \partial k = \lambda [ - (\partial P_h / \partial k) H + \partial T / \partial k ] = 0$$

$$(8) \quad \partial M / \partial \lambda = 24w + y - [P_h - P_h(k)]H - P_X X - wL - T - T(k, w) = 0$$

If one sets  $P_X = 1$  and eliminates  $\lambda$ , then he can rewrite conditions (4)-(7) as:

$$(9) \quad U_H / U_X = P_h - P_h(k)$$

$$(10) \quad U_L / U_X = w$$

$$(11) \quad - (\partial P_h / \partial k) H + \partial T / \partial k = 0$$

Condition (9) states that the marginal rate of substitution between the composite good,  $X$ , and housing equals the price of housing space, and condition (10) asserts that the marginal rate of substitution between the composite good,  $X$ , and leisure time equals the wage rate. However, condition (11) indicates that the reduction in expenditure on housing resultant from a small increment in commuting distance equals the increase in expenditure on commuting thereby incurred.

To determine the effect of an increase in wage income on commuting distance one

\* Assistant professor, University of Oklahoma.

$$(12) \quad \begin{bmatrix} U_{XX} & U_{XH} & U_{XL} & 0 & -P_X \\ U_{HX} & U_{HH} & U_{HL} & \lambda \frac{\partial P_h}{\partial k} & -[P_h - P_h(k)] \\ U_{LX} & U_{LH} & U_{LL} & 0 & -w \\ 0 & \lambda \frac{\partial P_h}{\partial k} & 0 & \lambda \left[ \left( \frac{\partial^2 P_h}{\partial k^2} \right) H & 0 \right. \\ & & & & \left. - \partial^2 T / \partial k^2 \right] \\ -P_X & -[P_h - P_h(k)] & -w & 0 & 0 \end{bmatrix} \begin{bmatrix} dX \\ dH \\ dL \\ dk \\ d\lambda \end{bmatrix} = \begin{bmatrix} \lambda dP_X \\ \lambda dP_h \\ \lambda dw \\ \lambda (\partial^2 T / \partial k \partial w) dw \\ X dP_X + H dP_h + dT - dy \\ - \left( 24 - L - \frac{\partial T}{\partial w} \right) dw \end{bmatrix}$$

$$(13) \quad \frac{dk}{dw} = \frac{\lambda D_{34} + \left( \frac{\partial^2 T}{\partial k \partial w} \right) \lambda D_{44} - \left( 24 - L - \frac{\partial T}{\partial w} \right) D_{54}}{D}$$

should totally differentiate conditions (4)–(8) which yields equation system (12). The next step is to solve system (12) for the change in commuting distance due to a change in wage income by using Cramer's rule. This gives  $dk/dw$  as shown in equation (13).

A change in the wage rate exerts three distinct effects on commuting distance. The first term in equation (13) is the compensated cross price effect between leisure and commuting distance. It can either be positive, which denotes that leisure and commuting distance are substitutes, or it can be negative, which means that leisure and commuting distance are complements. The second term in (13) is the compensated own price effect for commuting distance weighted by  $(\partial^2 T / \partial k \partial w)$ . It must be negative provided that the second-order conditions for a constrained utility maximum are satisfied. The third term in equation (13) is the Slutsky income effect for commuting distance multiplied by  $(24 - L - \partial T / \partial w)$ . It can either be positive, which signifies that commuting distance is a superior good, or it can be negative which indicates that commuting distance is an inferior good.

Equation (13) is not the same as Muth's formula for the effect of a change in wage income on commuting distance. He finds that a change in wage income will increase commuting distance only if the income elasticity of demand for housing space exceeds unity. Whereas I find that a rise in wage income can increase commuting distance only if the

sum of the three terms in equation (13) is positive.

In particular, the compensated cross price effect between leisure and commuting distance which appears in equation (13) is absent from Muth's result. Thus, if leisure and commuting distance are good substitutes my equation (13) is more likely to show that a rise in wages will increase commuting distance. However, if leisure and commuting distance are good complements this would not alter Muth's result because the compensated cross price effect between leisure and commuting distance doesn't enter into his result.

In summary, when leisure and commuting distance are independent goods, equation (13) and Muth's result yield similar conclusions about the effect of a change in wage income on commuting distance. But, when leisure and commuting distance are good substitutes, or good complements, equation (13) and Muth's result yield divergent findings for the effect of a change in wage income on commuting distance. In the former case, equation (13) is more likely than Muth's result to predict that a rise in wage income will tend to increase commuting distance, and in the latter case, equation (13) is more likely than Muth's result to predict that a rise in wage income will tend to reduce commuting distance.

#### REFERENCES

- J. Henderson and R. Quandt, *Microeconomic Theory*, New York 1958.  
R. Muth, *Cities and Housing*, Chicago 1969.

# Choice Involving Unwanted Risky Events and Optimal Insurance

By J. M. PARKIN AND S. Y. WU\*

This paper examines the consumer's choice involving insurance and gambling when he is faced with an unwanted contingency where its occurrence is uncertain. When the true state of the world is not known, an action taken by the decision maker may not yield a unique consequence. Under these circumstances, economists postulate that the decision maker will choose an action which maximizes his expected utility derived from the utility function defined over an appropriate domain of consequences. Traditionally, economists, following either John von Neumann and Oskar Morgenstern or Leonard Savage, assume that the decision maker evaluates the possible consequence of an action without considering the state of the world that obtains. In this case, the expected utility is calculated from the utility function defined over a set of what we may call "ordinary consequences." Let  $S$  denote the set of states of the world,  $X$  the set of ordinary consequences, and  $F$  the set of available acts. A weak order is assumed on  $F$ . This weak order induces a utility function  $u$  defined on  $X$  and a probability measure  $P$  on the set of all subsets of  $S$ . The expected utility is

$$(1) \quad E[u(f(s)), P] = \sum_{i=1}^n u(x_i) P(f = x_i)$$

where  $f \in F$ ,  $s \in S$ , and  $x_i \in X$ .

A wide class of choice problems involves consequences the enjoyment of which varies with the state that obtains. For example, a

young man will propose marriage to a young lady who may or may not accept. Surely his preference ordering over a commodity set (the ordinary consequences) is dependent upon whether he is married or single. Or a consumer who faces the possibility of illness must have a different preference ordering over a commodity set depending upon whether he is sick or well. Examples of this nature are abundant, all of which may be considered as involving an *ex post* shift in the consumer's utility function whenever the contingency occurs. But a choice problem involving an unwanted contingency is by its nature an *ex ante* process. For this class of uncertainty problems, the expected utility is calculated from the utility function defined over the set of "state-augmented consequences." Let  $A, B$  be disjoint subsets of  $S$ ,  $f_A$  and  $g_B$  be actions on  $A$  and  $B$ , respectively, where  $f_A$  and  $g_B$  are elements of a set  $D$ , and let  $C$  be the set of state-augmented consequences where  $C = S \times X$ . A weak order is assumed on  $D$ . R. Duncan Luce and David Krantz have shown that when certain structural and behavioral axioms are satisfied, there emerges a utility function  $w$  on  $C$  and a probability measure  $P$  on subsets of  $S$ . The expected utility can then be represented by

$$(2) \quad E(w, P) = w(x, A)P(A | A \cup B) + w(x, B)P(B | A \cup B)$$

where  $(x, \cdot) \in C$ . Note in general that  $w(x, A) \neq w(x, B)$ .

The treatment of the class of uncertainty problems described here is relatively unfamiliar in the economic literature. The pioneering work is associated with Jack Hirschleifer (1966). This relative scarcity is due perhaps to the lack of a theoretical (axiomatical) foundation which became available only recently through the works of Luce and Krantz.

\* Professors of economics, University of Manchester, England, and University of Iowa, respectively. This research was undertaken during Wu's tenure as a Special Fellow of the U.S. National Institutes of Health (HS 46293-01). The views expressed in this paper are our own, and do not reflect those of any branch or agency of the U.S. Public Health Service. We wish to thank Camilo Dagum, Joseph Stiglitz, Roy Ruffin, Hajime Hori, and especially Michael Balch for their many helpful comments.

In this paper, we employ the Luce-Krantz expected utility setup to analyze the consumer's choice involving insurance and gambling when he is faced with an unwanted contingency. Milton Friedman and Savage, in their celebrated article, have examined the influence of risk on consumer's choice with respect to insurance and gambling on the basis of the von Neumann-Morgenstern expected utility setup. In doing so, they restricted the alternatives open to the consumer in terms of income alone and assumed that the consumer's income function and tastes are kept invariant across states of the world. They then concluded that if the consumer's utility of income function is concave in the range of the uncertain income alternatives, he will choose to insure; that if the function is convex in this range, he will choose to gamble. Under the Friedman-Savage framework, significant contributions have been made in the areas of pure theory of choice (see Kenneth Arrow (1965) and John Pratt), portfolio analysis (see David Cass and Joseph Stiglitz, and James Tobin), insurance theory (Arrow (1963) and Karl Borch), welfare economics (Jerome Rothenberg), etc. The restrictive nature of the Friedman-Savage framework is evident. It not only failed to deal with the problems where the consumer's utility is dependent upon the state that obtains, but by restricting the choice alternatives to income alone, it also prevents the analysis of problems involving changes in relative prices which yield a substitution effect. By using the Luce-Krantz expected utility, we are able to relax the restrictions imposed by Friedman and Savage and analyze a wider class of choice problems involving insurance and gambling. As will be shown, the Friedman-Savage result becomes a special case of ours. In addition, under the proposed framework, commodities instead of income are the arguments of the utility function and the effect of a change in relative prices on the demand for insurance can easily be examined.

### I. The Model

In the proposed model, the arguments of the utility function are the elements of a

commodity vector rather than income; the preference function is allowed to be different depending upon whether the risky event does or does not occur. The criterion function to be maximized is the expected utility of the states of the world.

Let there be two states of the world,  $S = \{0, 1\}$ ; here  $A = \{0\}$  and  $B = \{1\}$ . The consumer's utility function is  $w(x, \alpha)$ , where  $x$  is a vector of commodities and  $\alpha = 0$  or  $1$ . For notational simplicity, let  $w(x, 0) \equiv U(x^0)$ ,  $w(x, 1) \equiv V(x^1)$ ,  $\pi \equiv P(A|A \cup B)$  and  $(1 - \pi) \equiv P(B|A \cup B)$ , where  $0 \leq \pi \leq 1$ . The Luce-Krantz expected utility becomes

$$(3) \quad E(\cdot) = \pi U(x^0) + (1 - \pi)V(x^1)$$

Not only does the utility function depend upon the state of the world but the constraint does also. We assume that the consumer may purchase an insurance policy for  $q$  dollars which will pay  $I(q)$  dollars in the event of  $\alpha = 1$ . We shall refer to  $q$  as the premium. The constraint associated with  $\alpha = 0$  is

$$(4) \quad y - q = p^T x^0$$

where  $y$  denotes the consumer's money income and  $p^T$  is a row vector of prices. The constraint in the event of  $\alpha = 1$  is

$$(5) \quad y' - q + I(q) = p^T x^1, \quad \text{where } y' < y$$

The problem facing the consumer is to choose  $x$  and  $q$  such that (3) is maximized subject to the constraints (4) and (5). It is intuitively appealing to view this maximization problem in two stages: First, choose  $x$  conditional upon  $\alpha$ , given  $q$ . Second, choose  $q$  using the information derived from the first stage.<sup>1</sup>

The first-stage problem is the conventional consumer choice problem and yields first-order conditions of the form<sup>2</sup>

$$(6) \quad \begin{aligned} &\text{for } \alpha = 0: \\ &U_{x^0} = \lambda p \\ &y - q = p^T x^0 \end{aligned}$$

<sup>1</sup> This two-stage maximization procedure although not widely used in economics is frequently used in statistics under the title of preposterior analysis.

<sup>2</sup> For simplicity, in this paper we assume that equality holds for  $U_{x^1} = \lambda p$  and  $V_{x^1} = \mu p$ , even if certain commodity  $x_i$  is not demanded.

(7) for  $\alpha = 1$ :

$$\begin{aligned} V_{x^1} &= \mu p \\ y' - q + I(q) &= p^T x^1 \end{aligned}$$

where  $\lambda$  and  $\mu$  are Lagrangean multipliers. Assuming the second-order condition to be satisfied, the first-order conditions can be solved to give demand functions of the form:

$$(8) \quad x^0 = F(p, y - q)$$

$$(9) \quad x^1 = G[p, y' - q + I(q)]$$

where  $x^0$ ,  $x^1$ ,  $F$ , and  $G$  are vectors.

The second-stage problem is to choose the optimal insurance premium  $q$  to maximize  $E(\cdot)$ . This is achieved by substituting the demand functions (8) and (9) into the utility functions in (3) yielding:

$$(10) \quad E(\cdot) = \pi U[F(p, y - q)] + (1 - \pi) V[G(p, y' - q - I(q))]$$

The consumer chooses the optimum value of  $q$  such that  $E$  is maximized by satisfying the conditions:  $dE/dq = 0$  and  $d^2E/dq^2 < 0$ .

The first-order condition requires that

$$(11) \quad -\pi U_{x^0}^T F_2 + (1 - \pi) V_{x^1}^T G_2 (I_q - 1) = 0$$

where  $T$  denotes transposition and the subscript denotes the argument with respect to which the derivative is taken. Equation (11) can be simplified by using the Engel aggregation condition. From the first-order conditions (6) and (7) we know that

$$U_{x^0} = \lambda p \quad \text{for } \alpha = 0$$

$$V_{x^1} = \mu p \quad \text{for } \alpha = 1$$

From Engel aggregations we know that

$$p^T F_2 = 1 \quad \text{for } \alpha = 0$$

and

$$p^T G_2 = 1 \quad \text{for } \alpha = 1$$

Hence

$$U_{x^0}^T F_2 = \lambda$$

and

$$V_{x^1}^T G_2 = \mu$$

Equation (11) becomes

$$(12) \quad \pi \lambda + (1 - \pi) \mu = (1 - \pi) I_q \mu$$

Notice that  $\lambda$  and  $\mu$  have their usual interpretation of marginal utility of net income. Equation (12) thus can be interpreted as follows: the consumer, when spending an extra dollar for insurance premiums, will forego expected utility of an amount  $[\pi \lambda + (1 - \pi) \mu]$  and will gain expected utility of an amount  $(1 - \pi) I_q \mu$ . The amount of insurance purchased is optimal when the expected marginal utility foregone is equal to the expected marginal utility gained.

Solving equation (12) for  $I_q$ , we obtain

$$(13) \quad I_q = 1 + \frac{\pi}{1 - \pi} \frac{\lambda}{\mu}$$

Suppose the  $I(q)$  function is linear. Then  $I_q$  is a constant. There are two classes of possibilities. First, the insurance scheme may be actuarial with no administrative costs. Second, the scheme may be nonactuarial. In the actuarial case,  $I(q)$  is given by:

$$(14) \quad I(q) = \frac{1}{1 - \pi} q$$

Therefore,

$$I_q = \frac{1}{1 - \pi}$$

From equation (13), this implies that  $\lambda = \mu$ . In words, when insurance scheme is actuarial, optimality requires that the marginal utility of net income be equalized in both states of the world.

Suppose the insurance scheme is not actuarial, but takes the form

$$(15) \quad I(q) = \frac{1}{1 - \pi} (q - k)$$

Then clearly, the optimality conditions between the actuarial and nonactuarial schemes are identical. However, if the insurance scheme takes the form<sup>3</sup>

$$(16) \quad I(q) = \beta \left( \frac{1}{1 - \pi} \right) q,$$

<sup>3</sup> In case the government or the employer contributes to the payment of the individual's premium, it is possible that  $\beta > 1$ . We ignore this possibility in this paper.

where  $\beta$  is a constant and  $1-\pi < \beta < 1$ , then

$$I_q = \frac{\beta}{1-\pi}$$

which implies that at the optimal point

$$(17) \quad \frac{\lambda}{\mu} < 1, \quad \text{hence } \lambda < \mu$$

In this nonactuarial case, optimality requires that the marginal utility of net income be less in state  $\alpha=0$  than in state  $\alpha=1$ . In other words, the optimum insurance in this case will be less than that of an actuarial scheme.

A further possibility is that the consumer's subjective probability ( $\pi^*$ ) of  $\alpha=0$  is different from the actuarial probability ( $\pi$ ). In this case, the first-order conditions will give

$$(18) \quad I_q = 1 + \frac{\pi^*}{1-\pi^*} \frac{\lambda}{\mu}$$

which must be equal to  $\beta/(1-\pi)$ , where the scheme is actuarial if  $\beta=1$  and is nonactuarial if  $1-\pi < \beta < 1$ .

Considering first the case  $\beta=1$ , putting (18) equal to  $1/(1-\pi)$ ,

$$(19) \quad \frac{\lambda}{\mu} = \frac{\pi}{\pi^*} \frac{1-\pi^*}{1-\pi}$$

which implies  $\lambda > \mu$  if  $\pi > \pi^*$  and  $\lambda < \mu$  if  $\pi < \pi^*$ . This result says that optimality requires the consumer to overinsure if he is a pessimist ( $\pi^* < \pi$ ) and underinsure if he is an optimist ( $\pi^* > \pi$ ).

For the case  $1-\pi < \beta < 1$ , we have

$$(20) \quad \frac{\lambda}{\mu} = \frac{1-\pi^*}{1-\pi} \frac{\pi}{\pi^*} - \frac{1-\pi^*}{1-\pi} \frac{1-\beta}{\pi^*}$$

The second part of this expression is positive. Hence, there will be a smaller tendency to overinsure in the case of pessimism ( $\pi^* < \pi$ ) and a greater tendency to underinsure in the case of optimism ( $\pi^* > \pi$ ).

The analysis has proceeded thus far on the assumption that the second-order conditions are satisfied. Now we must examine these conditions. From equation (12) we obtain the second-order conditions as:

$$(21) \quad \frac{d^2 E}{dq^2} = \pi \lambda_2 + (1-\pi) \mu_2 (I_q - 1)^2 + (1-\pi) \mu I_{qq} < 0$$

Using the first-order condition solution for  $I_q$  and assuming  $I_{qq}=0$ , (21) can be simplified<sup>4</sup> to

$$(22) \quad \frac{\lambda_2}{\lambda} < (1-I_q) \frac{\mu_2}{\mu}$$

Notice that, since  $\lambda$  and  $\mu$  are marginal utilities of income, the expressions  $\lambda_2/\lambda$  and  $\mu_2/\mu$  are the negative of the Arrow's absolute and Pratt's relative risk aversion coefficients,<sup>5</sup> denoted by  $r(\lambda)$  and  $r(\mu)$ . Then, (22) becomes

$$(23) \quad r(\lambda) > (1-I_q) r(\mu)$$

Since  $I_q > 1$  would be necessary for any insurance to be optimal, it is clear that  $(1-I_q) < 0$ . Hence, if both the utility of income functions are concave, the second-order condition is satisfied and insurance will be purchased. Also, it is clear that if both the utility of income functions are convex, the second-order condition is violated; then a solution of no-insurance becomes optimal. It is interesting, however, to examine what happens when one of the utility of income functions is convex and the other concave. In this case, it is still possible that insurance will be purchased. If  $r(\lambda) < 0$ , and  $r(\mu) > 0$ , i.e., if the  $\lambda$  function is convex and the  $\mu$  function is concave, then the consumer may still buy insurance if

$$(24) \quad -\frac{r(\lambda)}{r(\mu)} < (I_q - 1)$$

On the other hand, if  $r(\mu) < 0$  and  $r(\lambda) > 0$ , then the condition for some nonzero insurance to be optimal is

$$(25) \quad -\frac{r(\lambda)}{r(\mu)} > (I_q - 1)$$

In both cases, for the consumer to purchase

<sup>4</sup> When  $I_{qq} \neq 0$ , this condition becomes  $(\lambda_2/\lambda) < (1-\frac{1}{2}I_q)(\mu_2/\mu)$ . Since the second-order condition for  $I_{qq} \neq 0$  is not theoretically different from the linear case, we omit the discussion of it in the text.

<sup>5</sup> See K. J. Arrow (1965), p. 33 and Pratt, p. 125.

insurance, a sufficiently high risk aversion coefficient, *ceteris paribus*, is required.

To give (23) a more natural interpretation, let us assume that the two states of the world be health and sickness: An individual who is a risk averter, regardless of the state of his health, will always insure; one who is a risk seeker, in both states, will never insure. However, a person who is a risk seeker when healthy but a risk averter when ill by (24) (or the reverse (25)) may insure depending upon the relative strength of his risk attitudes in the two states and on the net marginal return from the insurance scheme.

Notice that the traditional analysis of Friedman-Savage which assumes that the utility function is independent of the state of world, can be viewed as a special case of the proposed model. This would imply that  $w(x, 0) \equiv w(x, 1)$  or  $U(x) \equiv V(x)$ . Then  $\lambda$  and  $\mu$  and their derivatives will be evaluated at different points on the same function. The comparison can best be demonstrated by using the example of an actuarial insurance scheme. For this case we have established that optimality requires that  $\lambda = \mu$ . The second-order condition represented by (21) is reduced to

$$(26) \quad \lambda_2 < \mu_2(1 - I_q)$$

Since  $U(x) = V(x)$ , the condition  $\lambda = \mu$  implies that  $\lambda_2 = \mu_2$ . But since  $I_q > 0$  always,  $\lambda_2 < 0$  is therefore required to satisfy (26). This is precisely the Friedman-Savage result.

## II. Effect of Price Change

So far we have held the commodity prices as given and examined the consumer's choice of insurance involving a change in the preference function. Now let us turn to examine the effect of changes in commodity prices on the demand for insurance. Assume the price of commodity  $i$  alone has changed. From (12), we obtain

$$(27) \quad \frac{dq}{dp_i} = \frac{\frac{1}{\mu} \frac{\partial \mu}{\partial p_i} - \frac{1}{\lambda} \frac{\partial \lambda}{\partial p_i}}{(1 - I_q) \frac{\mu_2}{\mu} - \frac{\lambda_2}{\lambda} - I_{qq}/(I_q - 1)}$$

A change in the price of the  $i$ th commodity will affect the marginal utility of income in both states of the world through two avenues: the first is the price change itself, which will be referred to as the price-effect, and the second is the induced change in the quantity of insurance purchased, which will be referred to as the insurance-effect. The price-effect and the insurance-effect manifest themselves, respectively, in the numerator and denominator of (27).

The second-order condition represented by the inequality (21) requires that the denominator of (27) be positive; i.e., the insurance-effect be positive. Therefore,

$$(28) \quad \frac{dq}{dp_i} > 0 \quad \text{if} \quad \frac{1}{\mu} \frac{\partial \mu}{\partial p_i} > \frac{1}{\lambda} \frac{\partial \lambda}{\partial p_i}$$

The price-effect may produce different percentage changes in the marginal utility of net income between the two states. A smaller (larger) amount of insurance will be purchased following a reduction of  $p_i$  if the price reduction causes a greater (smaller) percentage change in the marginal utility of net income for the state  $\alpha = 1$  than for the state  $\alpha = 0$ . In other words, when a price reduction causes a relatively greater percentage reduction in the marginal utility of money for the state  $\alpha = 1$ , in order to restore equilibrium, the consumer will reduce the amount of insurance purchased.

## REFERENCES

- K. J. Arrow, *Aspects of the Theory of Risk-Bearing*, Helsinki 1965.
- , "Uncertainty and the Welfare Economics of Medical Care," *Amer. Econ. Rev.*, Dec. 1963, 53, 941-73.
- K. Borch, "Recent Developments in Economic Theory and Their Application to Insurance," *ASTIN Bulletin*, Apr. 1963, 2, 322-41.
- D. Cass and J. Stiglitz, "The Structure of Investor Preferences and Asset Returns, and Separability in Portfolio Allocation: A Contribution to the Pure Theory of Mutual Funds," *J. Econ. Theor.*, June 1970, 2, 122-60.
- P. Fishburn, *Utility Theory for Decision Making*, New York 1970.

- M. Friedman and L. J. Savage, "The Utility Analysis of Choices Involving Risk," *J. Polit. Econ.*, Aug. 1948, 56, 279-304.
- J. Hirschleifer, "Investment Decisions Under Uncertainty-Choice-Theoretic Approaches," *Quart. J. Econ.*, Nov. 1965, 79, 509-36.
- , "Investment Decision Under Uncertainty: Applications of the State Preference Approach," *Quart. J. Econ.*, May 1966, 80, 252-77.
- R. D. Luce and D. H. Krantz, "Conditional Expected Utility," *Econometrica*, Mar. 1971, 39, 253-271.
- J. W. Pratt, "Risk Aversion in the Small and in the Large," *Econometrica*, Jan.-Apr. 1964, 32, 122-36.
- J. Rothenberg, *The Measurement of Social Welfare*, Englewood Cliffs 1961.
- L. J. Savage, *The Foundations of Statistics*, New York 1954.
- J. Tobin, "Liquidity Preference as Behavior Toward Risk," *Rev. Econ. Stud.*, Feb. 1958, 25, 65-86.
- J. von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*, Princeton 1947.

# A Spectral Analysis of Post-Accord Federal Open Market Operations: Comment

By LLAD PHILLIPS AND ROBERT WEINTRAUB\*

In a recent issue of this *Review*, Vittorio Bonomo and Charles Schotta (B-S) presented a model and evidence on Federal Reserve (FR) defensive open market operations. Their paper is seriously flawed. This paper calls attention to their errors and presents a correct application and testing of their model.

## I. The B-S Model and Results

B-S decomposed the sources of week-to-week changes in reserves, which they called  $R$ , into changes in Federal Reserve holdings of U.S. government securities, called  $G$ , and the catchall series  $R$  net of  $G$ , which they denoted by  $R^*$ . Algebraically,

$$(1) \quad R_t \equiv G_t + R_t^*$$

They asserted that  $G_t$  is a function of  $R_t^*$  but that  $R_t^*$  is in no way a function of  $G_t$  (see p. 51). From this identity it of course follows that the cycles in  $R^*$  will generate cycles in  $R$  unless offset by  $G$ . Investigating the period April 4, 1951–May 31, 1967 as a whole, B-S found 1) monthly cycles in  $R^*$  and  $G$ , 2) no monthly peak in  $R$ , and 3) no cross covariance between  $G$  and  $R$ .<sup>1</sup> Based on these results they concluded "... that the effect of Federal Reserve open market operations since the Accord has been practically to eliminate the very strong monthly element cycling in member bank reserves, ..." (p. 59). In addition, they found a monthly cycle in currency outside all banks and significant

coherence between this series and both  $R^*$  and  $G$ . On the basis of these data they suggested "... an interpretation of open market operations as being undertaken, or at least having the effect of existing primarily, for the purpose of offsetting a monthly cycle in currency outside all banks" (p. 60). But B-S committed important methodological and analytical errors which require us to set aside their results and conclusions and search for correct ones.

## II. The B-S Errors

From the methodological standpoint, B-S erred in a fundamental sense by failing to take into account that the definitions of  $R$  and  $R^*$ , two of the three variables in their model, were changed by legislation in 1959 permitting vault cash to be counted as reserves. Before the law, the series on week-to-week changes in member banks' vault cash was included in  $R^*$  but not  $R$ ; afterwards it was included in  $R$  but not  $R^*$ . It is bad procedure to analyze a time-series which measures one thing for the first half of the study period and another thing for the second half. The B-S evidence on  $R$  and  $R^*$  is just not credible because they did exactly this. Later, we shall introduce proper evidence.

A second methodological error committed by B-S was to ignore contributions to their catchall series  $R^*$  other than the one made by week-to-week changes in currency outside all banks.<sup>2</sup> Most importantly, they failed to in-

\* University of California at Santa Barbara.

<sup>1</sup> We confine our comments to the B-S results and conclusions on the monthly cycles in  $R$  and the factors that contribute to  $R$ . B-S also discussed findings on cycles at the 2.2 week frequency and presented and discussed evidence on cycles in free reserves, called  $FR$ . Our criticisms of the B-S methodology, though made in the context of their discussion of the monthly cycles in  $R$  and its components, apply to these other discussions as well.

<sup>2</sup> Furthermore, the B-S series on week-to-week changes in currency outside all banks does not measure the currency changes which in principle affect Federal Reserve defensive operations. This follows from the consideration that changes in currency outside all banks may reflect transfers of currency between the public and nonmember banks. Insofar as a cycle in currency outside all banks reflects one in nonmember bank vault cash, there is no effect on member banks and no reason for the Fed to respond. Clearly currency outside member banks only is the proper index of the currency changes which in principle influence the Fed's behavior.

investigate the contribution of changes in float and compare it with that of currency.

From the theoretical standpoint, B-S erred in two ways. First, they failed to interpret their model to cover the implications of the vault cash law. Second, they assumed away possible feedback from the behavior of  $G$  to  $R^*$ . Since  $R^*$  includes week-to-week changes in discounts and advances ( $DA$ ), and  $DA$  and  $G$  are potential substitutes for smoothing  $R$ , this assumption is clearly fallacious. As discussed below, these errors caused B-S to misinterpret their statistical results and prevented them from formulating and testing several interesting hypotheses including one on the role of  $DA$ .

### III. The B-S Model Interpreted to Cover the Vault Cash Law

The B-S model consists of the equation (1). To interpret this model to cover the vault cash law it is necessary to recognize that in the years since vault cash has been used to satisfy reserve requirements the following definitions are valid:

$$(2) \quad R \equiv MDF + VC$$

$$(3) \quad R^* \equiv [DA + F + OTH - COM]$$

When vault cash was not counted as reserves<sup>3</sup>, however, we had:

<sup>3</sup>  $R^*$  is derived by rearranging the factors that supply and absorb reserves. Week-to-week changes in the supply of reserve funds equal changes in the Fed's holdings of government securities,  $G$ , in discounts and advances,  $DA$ , in float,  $F$ , in the gold stock,  $g$ , and in treasury currency outstanding,  $tco$ . Week-to-week changes in the absorption of reserve funds equal changes in total currency and coin outstanding,  $CC$ , in treasury cash holdings,  $tch$ , in deposits other than member bank reserves with  $FR$  banks,  $od$ , in other  $FR$  accounts,  $oa$ , and in member bank deposits with  $FR$  banks,  $MDF$ . By the identity of the supply and absorption factors, we have  $G + DA + F + g + tco = CC + tch + od + oa + MDF$ . Letting  $OTH = g + tco - (tch + od + oa)$  and decomposing  $CC$  into the sum of changes in currency outside member banks,  $COM$ , and member bank vault cash,  $VC$ , we obtain  $G + DA + F + OTH = COM + VC + MDF$ . Transposing  $G$  and  $COM$ , we obtain  $[DA + F + OTH - COM] = (VC + MDF) - G$ . For the period when  $VC + MDF$  defined  $R$ ,  $[DA + F + OTH - COM]$  defines  $R^*$ . When  $MDF$  defined  $R$  then  $R^* = [DA + F + OTH - (COM + VC)]$ . For the record, we note also that since 1967 the series called  $oa$  has been split into one registering miscellaneous asset changes and another registering miscellaneous liability changes.

$$(4) \quad R \equiv MDF$$

$$(5) \quad R^* \equiv [DA + F + OTH - (COM + VC)]$$

The notations we introduce are,

$MDF$  = week-to-week changes in member bank deposits with  $FR$  banks.

$VC$  = week-to-week changes in member bank vault cash.

$DA$  = week-to-week changes in discounts and advances.

$F$  = week-to-week changes in float.

$OTH$  = week-to-week changes in the sum of sundry miscellaneous accounts: the gold stock + treasury currency outstanding - (treasury cash holding + deposits other than member bank reserves with  $FR$  banks + other  $FR$  accounts).<sup>4</sup>

$COM$  = week-to-week changes in currency outside member banks.<sup>5</sup>

### IV. Correcting the B-S Misinterpretations

B-S were remiss to interpret their findings on  $R$  as having any meaning whatever. Because the vault cash law changed the definitions of  $R$  and  $R^*$ , the periods before and after the law must be analyzed separately. We separated the period April 4, 1951 to November 25, 1959, when no vault cash counted as reserves, from the period November 30, 1960 to May 31, 1967, when all vault cash counted as reserves.<sup>6</sup> We found that the vault cash law was an important structural change. As shown in Table 1, before the vault cash law there was a significant monthly cycle in  $R$  and also significant coherence between  $G$  and  $R$ . After the law, there was no monthly cycle in  $R$ .

<sup>4</sup> Deposits other than member bank deposits with  $FR$  banks consist of three accounts. One of these, treasury deposits, exhibited small monthly cycles before and after the vault cash law. These were the only monthly cycles exhibited by any of the series that comprise  $OTH$ . We ignore them in the discussion because they are minor and because cross covariances involving this series were all insignificant.

<sup>5</sup> See fn. 2.

<sup>6</sup> Until November 25, 1959 no vault cash was counted as reserves. From December 1, 1959 to November 23, 1960 only part of member bank vault cash could be used to satisfy reserve requirements. From November 30, 1960 on all member bank vault cash counted as reserves.

TABLE 1—COMPARISON OF SPECTRAL AND CROSS-SPECTRAL ESTIMATES FOR THE PERIODS BEFORE AND AFTER THE VAULT CASH LAW<sup>a</sup>

Spectral Estimates at 4.4 Weeks				
Series	Percentage of Spectral Power in Each Series		Relative (Standardized) Spectral Peak Height <sup>b</sup>	
	Before	After	Before	After
<i>R</i>	10.6	3.0	2.30	.61 (insig.)
<i>G</i>	7.8	31.5	1.00	12.3
( <i>COM</i> + <i>VC</i> )	11.5	16.4	.85	1.76
<i>COM</i>	N.A. <sup>c</sup>	22.9	N.A.	4.97
<i>VC</i>	N.A.	15.7	N.A.	1.54
<i>DA</i>	15.2	2.8	1.70	.15 (insig.)
<i>FL</i>	25.2	22.8	6.15	7.70

Cross-Spectral Estimates <sup>d</sup>						
Series	Coherence		Gain		Phase Lag (in fraction of a circle)	
	Before	After	Before	After	Before	After
<i>R, G</i>	.37 <sup>e</sup>	.06 <sup>f</sup>	.41 ± .21		.30 ± .12	
<i>R, DA</i>	.65 <sup>e</sup>	.13 <sup>f</sup>	.70 ± .11		.38 ± .07	
<i>COM, VC</i>	N.A.	.99 <sup>e</sup>	N.A.	.55 ± .03	N.A.	.58 ± .01

<sup>a</sup> The period prior to the vault cash law is from the week ending April 4, 1951 to the week ending November 25, 1959. The period subsequent to the vault cash law commences with the week ending November 30, 1960 (at which date *all VC* was counted as *R*) and ends May 31, 1967.

<sup>b</sup> Relative to *G* (before the vault cash law) = 1.00. The 95 percent confidence intervals for the spectral power in the peaks lay above the confidence intervals for the neighboring troughs for all series reported unless noted as insignificant. Note also that peak height depends on total power as well as percentage of power.

<sup>c</sup> Series for *COM* and *VC* were not available prior to the vault cash law.

<sup>d</sup> 95 percent confidence intervals are indicated for phase lag and gain.

<sup>e</sup> Coherence square significantly different from zero at the 1 percent level.

<sup>f</sup> Coherence square not significantly different from zero at the 5 percent level.

Thus our results show that it was *wrong* for B-S to conclude that defensive open market operations had the effect of virtually eliminating a substantial monthly cycle in reserves in the period from April 4, 1951 to May 31, 1967. They did not have this effect in the early part of their study period, April 4, 1951–November 25, 1959, when vault cash was not counted as reserves.

B-S misinterpreted their findings on currency changes when they suggested that defensive open market operations were undertaken primarily to offset the monthly cycle they found in week-to-week changes in currency outside all banks, which they called *C*. This interpretation makes little sense when applied to the early period. The relationship

between their *C* and our *COM* series is given by<sup>7</sup>

$$(6) \quad C = COM - VCNM$$

where *VCNM* denotes week-to-week changes in nonmember banks' vault cash, and *COM*, as already stated, denotes week-to-week changes in currency outside member banks. Substituting (6) in (5) we obtain for the early period

$$(7) \quad R^* = [DA + F + OTH - (C + VCNM + VC)]$$

<sup>7</sup> The relationship between *C* and our index of the currency cycle, *COM*, follows from consideration that total currency outstanding, *CC*, equals both *COM*+*VC* and *C*+*VCNM*+*VC*.

The B-S hypothesis in question is that the monthly cycle in  $C$  generated one in  $R^*$  and thereby triggered defensive open market operations. But before the vault cash law there was little if any effect on  $R^*$  from  $C$ . This is because currency demands initially impact upon vault cash. Thus, before November 25, 1959 when the public withdrew currency from banks, the immediate and direct effect was that banks lost vault cash. In terms of (7), when  $C$  is positive  $VCNM$  and  $VC$  are negative and there is no change in  $R^*$ .

Of course it is *possible* that banks quickly replenish currency drains by exchanging deposits with  $FR$  banks for vault cash and absorb currency inflows by exchanging vault cash for deposits. In such a case the cycle in  $C$  would indirectly produce one in  $R^*$ . However, B-S didn't present this as the case. Moreover, there are strong reasons for rejecting it. As a matter of common sense economics, it would not have been economically feasible for the Fed to have permitted banks to exchange vault cash and  $FR$  deposits so as to replenish and absorb the monthly ebb and flow of currency, even if banks had wanted to. This is because there is a cost in transporting currency and coin. This cost is maximized if currency and coin transfers between  $FR$  banks and member banks are synchronized with the monthly currency cycle. Financial prudence dictates scheduling such transfers throughout the month, not at one or two time points. Add also that from an empirical standpoint there is no evidence that banks eliminated the cycle in vault cash which results directly from the currency cycle. Data are not available on  $VC$  or  $VCNM$  before the vault cash law. As shown in Table 1, after the law there was a significant monthly cycle in  $VC$ . (Data on  $VCNM$  are not available.) Also there was high coherence between  $VC$  and  $COM$  and the monthly cycles in  $VC$  and  $COM$  were out of phase. These data provide strong support for our contention that cost considerations prevent banks from exchanging vault cash and  $FR$  deposits to replenish and absorb vault cash changes deriving from currency cycles. It seems clear that a cycle in currency creates an offsetting one in vault cash. Hence we

reject the possibility that the monthly currency cycle, whether one chooses to index it by  $COM$  or  $C$ , indirectly produced one in  $R^*$  before 1960.

Thus, B-S were remiss to have interpreted the monthly cycle in currency as generating a monthly cycle in  $R^*$  and thereby being the underlying cause of defensive open market operations *throughout* their study period. Had they interpreted their model to cover the vault cash law, they might have seen as we have shown, that before the law  $R^*$  was unaffected by the cycle in  $C$  by reason that  $VC$  and  $VCNM$  also contributed to  $R^*$ , and vault cash is the direct absorbent of the currency cycle. Only since December 1960, that is since all vault cash has counted as reserves, has the currency cycle caused a cycle in  $R^*$  and thereby acted to trigger defensive open market operations.

#### V. Omitted Tests

Because they ruled out feedback from the behavior of  $G$  to  $R^*$  by assumption, B-S failed to explore the role played by discounting in smoothing  $R$ . As shown by identities (3) and (5),  $DA$  contributes to  $R^*$ . But  $DA$  is itself a control variable like  $G$ . Hence, in principle, if  $G$  is not doing the job of smoothing  $R$  by offsetting cycles in  $R^*$ , then,  $DA$  has a role to play. In this connection, as shown in Table 1, there was a significant monthly cycle in  $DA$  in the early period when  $R$  also exhibited a significant cycle, and moreover the coherence between  $R$  and  $DA$  was larger than between  $R$  and  $G$ .

Because they ignored factors contributing to  $R^*$  other than currency changes, B-S failed to explore the important role played by float. As shown in Table 1, there were significant large monthly cycles in  $F$  both before and after the vault cash law. Before the law the monthly cycle in  $F$  was the main contributor to the one in  $R^*$ , and hence provided the main reason for defensive discounting and open market operations. Since the law the cycle in  $COM$  also has played an important part in the process; for after the law currency changes impacted directly upon  $R^*$ .

Last, because they did not interpret their model to cover the vault cash law, B-S failed

to observe that the size of the monthly cycle in  $G$  was very much larger in the 1960's than in the 1950's. Observing this, we looked for reasons. The proximate reason for the substantial increase in the monthly cycle in  $G$  after the vault cash law was a substantial increase in the monthly cycle in  $R^*$ . In turn, the cycle in  $R^*$  rose for three reasons: 1) the peak height of the monthly cycle in  $F$  increased about 25 percent; 2) the monthly cycle in  $COM$  now impacted directly upon  $R^*$  since  $VC$  now was a component of  $R$ , and hence in the 1960's both the float and currency cycles had to be offset by control operations; and 3) after the vault cash law discounting was not used as a defensive control— $DA$  contained no significant power at the 4.4 week frequency, and hence the job of offsetting the monthly cycle in  $R^*$  fell entirely on open market operations.

Other questions remain unresolved. We have not explained why defensive operations failed to eliminate the monthly cycle in  $R$  before the vault cash law and why  $DA$  ex-

hibited a monthly cycle in the early period but not in the later period. These questions are not illuminated by the B-S model. To cast light on these questions requires specifying a behavioral structure which covers member bank desires. Also there remains the task of separating the changes in  $R$  which are produced by  $G$  from changes produced by other factors. This was a primary interest of B-S (see p. 51). But their model cannot cast light on this matter because it fails to deal with feedback from  $G$  to components of  $R^*$ . To treat this question a model must be specified which permits separating the exogenous elements of  $R^*$  from the endogenous ones. We are currently preparing research on the unresolved questions.

#### REFERENCE

- V. Bonomo and C. Schotta, "A Spectral Analysis of Post-Accord Federal Open Market Operations," *Amer. Econ. Rev.*, Mar. 1969, 59, 50-61.

# A Spectral Analysis of Post-Accord Federal Open Market Operations: Reply

By VITTORIO BONOMO AND CHARLES SCHOTTA\*

In the preceding comment by Llad Phillips and Robert Weintraub (P-W) on our article in this *Review*, a number of criticisms are advanced. In this reply, we shall show that these criticisms are either totally irrelevant to our article, wholly unsubstantiated, or merely replications of minor reservations that we had indicated about our own research findings.

Many of the critical statements arise simply from a fundamental misunderstanding as to the content of our article on the part of P-W. One important example should be sufficient to illustrate the nature of the misunderstanding.

In their comment, P-W state

Thus our results show that it was *wrong* of B-S to conclude that defensive open market operations had the effect of virtually eliminating a substantial monthly cycle in reserves in the period from April 4, 1951 to May 31, 1967. They did not have this effect in the early part of their study period, April 4, 1951–November 25, 1959, when vault cash was not counted as reserves.  
[p. 990]

In this statement, P-W appear to have conjured up a conclusion and attributed it to us; namely, the conclusion that the effects of open market operations (*OMO*) on the monthly cycle in reserves were *identical* over all possible segments of the post-Accord period, 1951–67. In our article, we nowhere say or even imply that the variance reduction effects of *OMO* on the monthly cycle in reserves were identical in each and every

year of the post-Accord period. We simply examined the data for the entire post-Accord period, reported our statistical findings for that period under study, and made it clear that our conclusions apply only to the *whole* period examined.

Contrary to P-W's assertion quoted above, *we were not wrong* to conclude that *OMO* greatly reduced the monthly cycle in reserves over the post-Accord period. Our statistical findings clearly support this conclusion and the reader may verify this by examining our article. P-W present no evidence whatsoever on the *variance reduction* effects of *OMO* on reserves for the *entire* sample period (1951–67) or, as a matter of fact, for any subperiod of the post-Accord period. And it is that sort of evidence that is necessary to refute our conclusion on the *variance reduction* effects of *OMO* for the entire post-Accord period. The spectral results that P-W present for two subperiods in no way vitiate our previous conclusion.

Keeping in mind the demonstrated tendency of P-W to have a flexible and highly imaginative view of the content of our article, we now turn our attention to what appears to be the major criticism advanced by P-W.

Phillips and Weintraub offer the testable hypothesis that legislation in 1959 permitting vault cash to be counted as bank reserves constituted an important structural change. Moreover they argue that this change casts doubt on our findings that *OMO* dampened the monthly cycle in reserves for the entire post-Accord period. They show that when vault cash is not included in reserves, changes in currency held by the nonmember bank public will have no effect on member bank reserves, *if* banks passively permit their vault cash positions to change with these currency flows, dollar for dollar. In such a case, a strong monthly cycle in currency in circulation would not produce or augment a

\* Associate professors of economics at Virginia Polytechnic Institute and State University. Schotta is currently on leave as economist, Research Division, Office of the Assistant Secretary for International Affairs, U.S. Treasury. Bonomo wishes to acknowledge financial support from the National Science Foundation, Grant No. 031487. We are indebted to Keith P. Russell for helpful comments.

monthly cycle in reserves, unless part or all of vault cash is counted as reserves.

After some discussion, P-W conclude that banks passively permit currency flows to alter their vault cash positions, dollar for dollar. They write that "It seems clear that a cycle in currency creates an offsetting one in vault cash. Hence, we reject the possibility that the monthly currency cycle . . . indirectly produced one in  $R^*$  before 1960" (p. 991). The reader will recall that  $R^*$  represents the change in member bank reserves arising from all other sources and uses of reserves *except* *OMO*. Thus, according to P-W, prior to the vault cash law, monthly fluctuations in currency outside member banks would affect only vault cash but not reserves. On the other hand, after the vault cash law, a monthly cycle in reserves would be produced or augmented by a monthly currency cycle, *assuming* that *OMO* did not fully offset the effects of currency flows.

Phillips and Weintraub then set out to determine statistically whether the vault cash law was an important structural change. The design of their test is as follows. If the calculated percent of the total spectral power in the monthly cycle frequency in reserves differs significantly for periods before and after the vault cash law, *and (or?)* if the coherence between  $OMO(\Delta G)$  and changes in reserves ( $\Delta R$ ) differs significantly over the two periods, then the vault cash law was an important structural change.

Utilizing spectral techniques, P-W investigate *weekly* data for two large subperiods of the post-Accord period. The pre-vault cash law period, 1951-1959, is approximately eight and one-half years in length; the post-vault cash law period, 1960-67, is about six and one-half years in length. Their statistical results are reported in their Table 1. On the basis of these results, their conclusion is that

We found that the vault cash law was an important structural change. As shown in Table 1, before the vault cash law there was a significant monthly cycle in  $R$  and also significant coherence between  $G$  and  $R$ . After the law there was no monthly cycle in  $R$ . [p. 989]

In our article (fn. 13, page 59), we stated that, at that time, we were also examining selected subperiods within the period 1931-1968 to ascertain (i) whether the dampening effects of *OMO* hold generally throughout the period and (ii) whether there is any evidence of an increasing ability or willingness over time on the part of the Fed to utilize *OMO* for short-run defensive purposes. The results of that study appeared in the June 1970 issue of the *Journal of Finance*.

P-W do not cite this subsequent evidence on defensive *OMO* nor do they give any indication that they are aware of its existence. It turns out, however, that the evidence presented in our 1970 article strongly suggests that P-W's breakdown of the post-Accord data period into two large subperiods is inappropriate for an adequate test of their "structural change" hypothesis.

In a segment of our 1970 study, we estimated an "offset coefficient" which we called  $\hat{k}_1$  which basically represents the estimated fraction of other factors affected reserves ( $\Delta R^*$ ) which tended to be offset by *OMO* for each year, 1931-68. By way of illustration, if  $\hat{k}_1 = -1$ , *OMO* are said to have had the effect of completely offsetting changes in other factors affecting reserves, thereby greatly reducing the short-run movements in reserves. Alternately, if  $\hat{k}_1 \approx 0$ , short-run defensive *OMO* are presumed to be nonexistent or completely ineffective.

In Table 1, we reproduce our estimates of  $\hat{k}_1$  for each of the years 1951-68. As an examination of Table 1 will confirm, there is a pronounced secular increase in the value of

TABLE 1

Year	$\hat{k}_1$	Year	$\hat{k}_1$
1951	-.08	1960	-.49
1952	-.05	1961	-.61
1953	-.17	1962	-.69
1954	-.12	1963	-.65
1955	-.35	1964	-.83
1956	-.37	1965	-.67
1957	-.47	1966	-.77
1958	-.52	1967	-.60
1959	-.32	1968	-.69

Source: Bonomo and Schotta (1970), Table 1, p. 662.

$k_1$  as we move from 1951 through 1968. For example, the average value of  $k_1$  for the period 1951-59 (which corresponds roughly to the pre-vault cash law period) is only  $-.27$ , whereas the average value for  $k_1$  since 1959 is approximately  $-.67$ .

The evidence in Table 1 clearly suggests that short-run fluctuations in reserves, including monthly cyclical fluctuations, are likely to have been much greater in the early part of the post-Accord period than in the middle part or latter part. To confirm this interpretation, we separately estimated the percent of total spectral power in reserves existing in the monthly cycle frequency for each of the years 1951 through 1968. We found that (i) the monthly cycle in reserves had an average value 68 percent higher in the period 1951-59 than in the period 1960-68, and (ii) the average value of the percent of spectral power residing in the monthly cycle for reserves declined continually over consecutive four-year subperiods of the post-Accord period.

As a result of this evidence on the increasing efficiency of defensive *OMO* since 1951, it should be obvious that one could select *any* historical event that occurred after, say, 1955 to be a candidate for an "important structural change." By comparing the spectral power in the monthly cycle (or total spectral power) in reserves for the early part of the post-Accord period (in this case, 1951 through 1955) with the spectral power obtained for the remaining part of the post-Accord period, one would, following the research methodology employed by Phillips and Weintraub, have to conclude that that historical event was an important structural change. That would mean that events such as President Eisenhower's reelection to a second term in 1956, and even Senator Goldwater's defeat for the presidency in 1964 were important structural changes which render the results in our article (1969) meaningless. Of course, this is nonsense. As economists working with time-series data, we must always strive to avoid the *post hoc, ergo propter hoc* fallacy.

Unless the secular influences are somehow removed or, at least, reduced, a simple com-

parison of spectral power for the *entire* period since 1951 prior to occurrence of a given historical event with the *entire* remaining segment of the post-Accord period will produce spurious results. It is impossible, in such a case, to know what effect, if any, the particular historical event had on the spectral power (and coherence between the variables) that was not properly due to other unspecified secular influences.

It should be noted that, following Phillips and Weintraub, the vault cash legislation should be expected to have *immediate* and rather *automatic* effects. This expectation follows from their *identity* equations and their strong assumption that banks are perfectly passive in their vault cash position behavior. Thus, the major effects of the vault cash law should be expected to be present in data corresponding to periods *immediately* before and after the vault cash law was instituted.

Since we (and P-W) are working with weekly data, it is reasonable to expect that an examination of three years of data before *and* after the vault cash law should be sufficient to determine the effects of the law, and, at the same time, keep the secular influences to a minimum.

We estimated the spectral and cross-spectral statistics for periods of one, two, and three years in length for both *before* the vault cash law and *after* the full vault cash inclusion in reserves. We present our spectral results in Table 2.

As an examination of Table 2 will confirm, the statistical results lend no support what-

TABLE 2

Number of Weekly Observations	Percent of Spectral Power in $\Delta R$ At 4.3 Week Cycle		Coherence { $\Delta G, \Delta R$ } At 4.3 Week Cycle	
	Before	After	Before	After
52	8.06	10.25	.25	.22
104	9.22	9.40	.14	.26
156	7.81	7.87	.19 <sup>a</sup>	.22 <sup>a</sup>

*Note:* For all periods, 17 frequencies were estimated.

<sup>a</sup> Coherence Value significantly greater than zero at the .95 probability level.

soever to P-W's hypothesis that the vault cash law was an important structural change, using the *same* criteria as P-W. P-W's results, derived from the *entire* post-Accord period, were that the percent of spectral power in reserves at the monthly cycle frequency was significantly greater before the law than the percent of spectral power in that frequency after the law. We find just the opposite! In every case, for 52, 104, and 156 observations before and after the law, the monthly cycle in reserves *after* the law exceeded the monthly cycle in reserves *before* the law. Hence, we cannot accept the P-W hypothesis that the vault cash law altered the percent of spectral power for reserves in the monthly cycle frequency.

With respect to the coherence statistics between *OMO* ( $\Delta G$ ) and reserves ( $\Delta R$ ), we found that coherence is not significantly different from zero for 52 and 104 observations for both the before and after periods. For 156 observations, *both* coherence statistics are significantly different from zero, with the numerical value of coherence for the *after* period being slightly larger. This is also a finding which is completely contrary to the finding by P-W. In their expanded-period results, the coherence between  $\Delta G$  and  $\Delta R$  was

significantly greater than zero *before* the law but not after.

Hence, we reject their assertion that the vault cash law was an important structural change. Consequently, whatever relevance this supposed structural change had to the findings in our article (1969), we reject all assertions by P-W that our conclusions for the post-Accord period were not correct.

And, finally, statements by P-W such as those which assert that we (i) "ruled out feedback from the behavior of  $G$  to  $R$ ," and (ii) "failed to explore the important role played by float" are merely replications of criticisms we had indicated about our own work. In rebuttal to statement (i) above, see footnote 4, p. 51 of our article and to statement (ii) above, see the concluding paragraph of our article (p. 60).

#### REFERENCES

- Y. Bonomo and C. Schotta, "A Spectral Analysis of Post-Accord Federal Open Market Operations," *Amer. Econ. Rev.*, Mar. 1969, 59, 50-61.
- , "Federal Open Market Operations and Variations in the Reserve Base," *J. Finance*, June 1970, 25, 659-67.

# Multi-Neutral Technical Progress: Compatibilities, Conditions, and Consistency With Some Evidence

By EARL R. BRUBAKER\*

In recent years great leaps forward have been made toward understanding logically possible and plausible hypotheses about the ways that technical progress might take place and toward assembling and analyzing evidence to test them. The nature of such hypotheses has been elaborated and generalized in an elegantly simple fashion by Martin Beckmann and Ryuzo Sato. Not content simply to state logical possibilities, they have assembled and analyzed data relating to a substantial segment of relevant, recorded human experience. Their analysis (1969, p. 92) hints that even further possibilities may be developed. Particular specifications might involve multi-neutrality, i.e., neutrality meeting some combination of single requirements such as Hicks-Harrod, Hicks-Solow, Hicks-Harrod-Solow, etc. Thus the number of logically possible hypotheses is seen to be proliferating, especially when one recalls that Beckmann and Sato have suggested six additional single specifications available for combination. On the surface it would appear that given nine single specifications we would have to consider a much larger number of possible combinations.

The proliferation of competing hypotheses may be seen to be even more unwieldy when we think about the precise algebraic specification of any single hypothesis. Thus, for example, the property of Hicks neutrality holds whenever the factor marginal productivity ratios are a function, *any* function, of the ratio of factor inputs alone. In the notation,

$$\frac{\partial Y/\partial K}{\partial Y/\partial L} = R\left(\frac{L}{K}\right)$$

$R(L/K)$  may be linear, *log-linear*, or any other function one may care to conjure up, and the implied production function would have the property of Hicks neutrality.

The purpose of the present paper is to explore some aspects of multi-neutrality by: 1) showing which of these are mutually compatible and which may be rejected on the grounds of logical inconsistency; 2) deriving the general conditions for compatible multi-neutralities; and 3) considering the consistency of the alternative hypotheses with some relevant evidence.

The analysis of the entire paper is based on certain important assumptions including: 1) efficiency in production or inefficiency impinging in like proportions on all inputs; and 2) a continuous, linearly homogeneous, single output, two input, production function with positive first partial derivatives and negative second partials.

The following notation will be employed:

$Y$  = output  
 $K$  = capital input  
 $L$  = labor input  
 $y = Y/K$   
 $x = L/K$   
 $z = Y/L$   
 $k = K/L$   
 $r$  = return to capital  
 $w$  = wage rate  
 $R = r/w$   
 $t$  = time

## I. Compatibility of Multi-Neutralities

The nine types of single neutrality, described respectively as Hicks, Harrod, Solow, Labor-combining, Capital-combining, Labor-decreasing, Capital-decreasing, Capital-additive, and Labor-additive, are generated

\* Associate professor of economics, University of Wisconsin, Madison. Comments by members of Professor Abram Bergson's Seminar on Comparative Economics at Harvard and by an anonymous referee have improved both exposition and substance of this paper. Support from the Graduate School of the University of Wisconsin and from the Russian Research Center at Harvard facilitated completion of the work.

from unique relationships between the following pairs of variables  $(r/w, L/K)$ ,  $(r, Y/K)$ ,  $(w, Y/L)$ ,  $(w, Y/K)$ ,  $(r, Y/L)$ ,  $(r/w, Y/K)$ ,  $(w/r, Y/L)$ ,  $(w, L/K)$ , and  $(r, K/L)$ .

On cursory examination it would appear that a large number of additional species of neutral technical change could be generated by requiring neutrality according to two or more of the criteria listed above. Fortunately, many combinations, in fact the great majority, are logically inconsistent. Table 1

TABLE 1—COMPATIBILITY OF MULTI-NEUTRALITIES<sup>a, b</sup>

Single Specification	I	II	III	IV	V	VI	VII	VIII
I $(r/w, L/K)$								
II $(r, Y/K)$	+							
III $(w, Y/L)$	+	+						
IV $(w, Y/K)$	-	-	-					
V $(r, Y/L)$	-	-	-	-				
VI $(r/w, Y/K)$	-	-	+	-	-			
VII $(w/r, Y/L)$	-	+	-	-	-	-		
VIII $(w, L/K)$	-	+	-	-	-	+	-	
IX $(r, K/L)$	-	-	+	-	-	-	+	-

<sup>a</sup> The letters  $r$  and  $w$  are used for notational convenience. The specifications of neutrality could be made in terms of  $\partial Y/\partial K$  and  $\partial Y/\partial L$  without assumption of their equality with  $r$  and  $w$ , respectively.

<sup>b</sup> Plus—Compatible; Minus—Incompatible.

summarizes the results of an investigation of compatibility of pairs from among the single varieties. As may be verified, twenty-seven of the possible thirty-six dual combinations are inconsistent. Systematic examination reveals that only one compatible triple combination can be found, i.e., the Hicks-Harrod-Solow combination, and there are no compatible combinations of four or more.

Thus instead of an overwhelmingly large number of competing hypotheses, the number may be reduced on logical grounds to only ten (nine dual combinations and one triple).

Compatibility may be determined through examination of the graph implied by the well-behaved production function in per capita terms. See Figure 1. Let us recall that in such a representation the relationship appears as a curve emanating from the origin with curvature concave toward the abscissa. The slope of a tangent to the curve measures the value

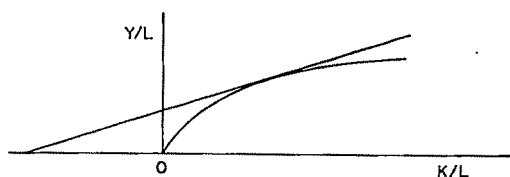


FIGURE 1

for  $\partial Y/\partial K$  associated with the values for  $Y/L$  and  $K/L$  at the point of tangency. The ordinate intercept measures  $\partial Y/\partial L$  and the abscissa intercept measures  $(\partial Y/\partial L)/(\partial Y/\partial K)$  (see R. G. D. Allen, p. 47). The slope of a ray from the origin to a point on the curve measures  $Y/K$ . Technical progress may be represented by an upward shift in the curve as time passes.

Using these concepts one can easily represent technical change that is neutral according to the various single specifications. Similarly from inspection of the graphs it is easy to determine if a pair of single specifications is compatible. Consider the examples presented with the aid of Figure 2.

First the pair,  $(w, L/K)$  and  $(r, Y/L)$ , is not compatible. The first requires that tangents from  $A$  and  $B$  have the same ordinate intercept. The second requires parallel tangents at  $A$  and  $C$ . Inspection of Figure 2 makes clear that if the first requirement is met, as is shown, the second cannot be.

The pair  $(w, L/K)$  and  $(r, Y/K)$  is compatible. Inspection of Figure 2 reveals that a common ordinate intercept of tangents from  $A$  and  $B$ , respectively, would not prevent the tangent at  $D$  from being parallel to that at  $A$  as required for Harrod neutrality. A complete analysis of the thirty-six possible pairs yields the results shown above in Table 1.

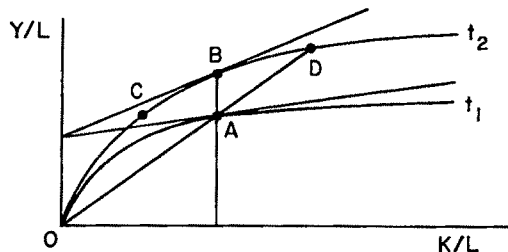


FIGURE 2

## II. Conditions for Multi-Neutral Production Functions

Beckmann and Sato have derived production functions implied by the various conditions for single neutrality. The question that immediately arises here is what production functions are implied by the various compatible multi-neutral combinations? Hirofumi Uzawa's proof that the Cobb-Douglas function is the only one both Harrod-neutral and Hicks-neutral is frequently mentioned by writers on technical change including, for example, Allen and C. E. Ferguson. Is there a function that meets the condition of Hicks, Harrod, and Solow neutrality? What of the other compatible dual neutralities? It is to a discussion of these questions that we now turn.

Each specification of single neutrality may be stated in the form of a differential equation in some combination of three variables,  $x$ ,  $y$ , and  $dy/dx$ .<sup>1</sup> "Solving" a combination of such specifications gives rise to a statement of a condition for the relations between variables sufficient for the existence of the properties of all the relevant single neutralities.

Thus, for example, Hicks, Harrod, and Solow neutrality require, respectively, that:

$$(1) \quad \frac{y - x \frac{dy}{dx}}{\frac{dy}{dx}} = R(x)$$

$$(2) \quad y - x \frac{dy}{dx} = g(y)$$

and

$$(3) \quad \frac{dy}{dx} = h\left(\frac{y}{x}\right)$$

Taking account of the three interrelationships<sup>2</sup> we find that:

$$(4) \quad \frac{g(y) + xh\left(\frac{y}{x}\right)}{dy} = \frac{R(x) + x}{dx}$$

<sup>1</sup> Recall that  $\partial Y/\partial K = y - (x)(dy/dx)$  and  $\partial Y/\partial L = dy/dx$  (see Allen, pp. 45-46).

<sup>2</sup> From (2) and (3) we see that  $y = g(y) + xh(y/x)$ . Substituting for  $y$  in (1) and rearranging terms gives (4).

Presumably we may choose  $g(y)$ ,  $h(y/x)$ , and  $R(x)$  in any way that we wish, and the production function obtained by solving the resulting differential equation will be neutral in the Hicks, Harrod, and Solow senses.

Where the specification of multi-neutrality involves a dual combination we have only two equations in three variables with an apparent possibility for obtaining three statements, each in two variables, of conditions sufficient for that dual neutrality. In many of these cases, however, the sufficient conditions may be stated in only two equations. Certain instances permit derivation of only one equation until after more precise specification of the single neutralities. These types of results may be illustrated as follows.

Let us consider first a familiar type of dual neutrality, the Hicks-Harrod variety. The single conditions are (1) and (2). Together they imply<sup>3</sup> that:

$$(5) \quad \frac{dy}{g(y)} - \frac{dx}{R(x)} = 0$$

or

$$(6) \quad yR(x) - xg(y) - g(y)R(x) = 0$$

Again it would appear that either of these conditions would suffice as a route to generation of a production function neutral in both (at least) the Hicks and the Harrod senses.

To illustrate a case in which more complete specification of a single neutrality is necessary, consider the Harrod-Solow variety of dual neutrality. Here the individual conditions are (2) and (3) which together imply

$$(7) \quad y = g(y) + xh\left(\frac{y}{x}\right)$$

For a general statement, (7) is as far as we can go. By specifying (3) more precisely, conditions additional to those covered by the

<sup>3</sup> Equations (1) and (2) may be written

$$(1a) \quad y - x \frac{dy}{dx} - \frac{dy}{dx} R(x) = 0$$

$$(2a) \quad y - x \frac{dy}{dx} - g(y) = 0$$

Subtracting and rearranging terms gives (5). Elimination of  $dy/dx$  gives (6).

TABLE 2—CONDITIONS FOR MULTI-NEUTRALITY

---



---

Hicks-Harrod-Solow:	$\frac{g(y) + xh(y)}{dy} - \frac{R(x) + X}{dx} = 0$
Hicks-Harrod	$\left\{ \begin{array}{l} \frac{dy}{g(y)} - \frac{dx}{R(x)} = 0 \text{ or} \\ yR(x) - xg(y) - g(y)R(x) = 0 \end{array} \right.$
Hicks-Solow <sup>a</sup>	
Harrod-Solow <sup>b</sup> :	$y - g(y) - xh\left(\frac{y}{x}\right) = 0$
Harrod-Capital-decreasing	$\left\{ \begin{array}{l} \frac{dy}{dx} - h\left(\frac{y}{x}\right)g(y) = 0 \text{ or} \\ h\left(\frac{y}{x}\right) - \frac{dy}{dx}\left(h\left(\frac{y}{x}\right)x + 1\right) = 0 \end{array} \right.$
Solow-Labor-decreasing <sup>a</sup>	
Harrod-Capital-additive <sup>b</sup>	$y - g(y) - xf(x) = 0$
Solow-Labor-additive <sup>a, b</sup>	
Labor-decreasing-Capital-additive <sup>b</sup> :	$y - h(x)g(y) - xh(x) = 0$
Capital-decreasing-Labor-additive	$\left\{ \begin{array}{l} \frac{dy}{dx} - h\left(\frac{y}{x}\right)g\left(\frac{1}{x}\right) = 0 \text{ or} \\ yxh\left(\frac{y}{x}\right) - \left(xh\left(\frac{y}{x}\right) + 1\right)\left(y + g\left(\frac{1}{x}\right)\right) = 0 \end{array} \right.$
Labor-decreasing-Capital-additive	

---

<sup>a</sup> From the symmetry of the roles of  $K$  and  $L$  in the production function it is clear that if we let  $y$  represent  $Y/L$  instead of  $Y/K$  and let  $x$  represent  $K/L$  instead of  $L/K$ , then  $dy/dx = \partial Y/\partial L$  instead of  $\partial Y/\partial K$ . Furthermore the configurations of symbols constituting statements of conditions for multi-neutrality in the Hicks-Harrod sense are transformed into a statement of multi-neutrality in the Hicks-Solow sense. A similar argument may be stated with respect to the following pairs of multineutralities: Harrod-Capital-decreasing and Solow-Labor-decreasing; Capital-decreasing-Labor-additive and Labor-decreasing-Capital-additive; Harrod-Capital additive and Solow-Labor-additive.

<sup>b</sup> Complete statement of the condition depends on specification of the wage as a function of  $Y/L$  where Solow neutrality is involved or as a function of  $L/K$  where Capital-additive neutrality is involved. See example in text.

general statement in (7) may be obtained. Suppose for example, we hypothesize that:

$$g(y) = ay \quad \text{and} \quad h\left(\frac{y}{x}\right) = b \frac{y}{x}$$

where  $a$  and  $b$  are parameters. Substitution in (7) tells us only that  $1 = a + b$ . If, however, the linear, zero-intercept specification is made in (3) at the outset, we can obtain<sup>4</sup> the more meaningful

$$(8) \quad \frac{dx}{x} \left( \frac{b}{1-b} \right) = \frac{dy}{g(y)}$$

<sup>4</sup> With  $dy/dx = (b)(y/x)$  by substituting for  $y$  in (2) we find

$$\frac{x \left( \frac{dy}{dx} \right)}{b} - x \left( \frac{dy}{dx} \right) = g(y)$$

which with a little further manipulation gives (8).

Table 2 presents a summary of the conditions for the various types of multi-neutrality. This summary is considerably shortened by reference to the symmetry of the roles of capital and labor in the production function. The meaning and implications of the symmetry are discussed in footnote a to Table 2.

One further comment might be made at this point. The conditions sufficient for neutrality or multi-neutrality do not guarantee that the production functions generated will be "well behaved."<sup>5</sup> Indeed in some cases what appear to be plausible specifications for single neutralities, e.g., a linear function with zero intercept, lead to "poorly behaved" functions in the context of multi-neutral technical progress.

<sup>5</sup> Well behaved or "good behavior" is used to indicate that over the range of possible values for  $K/L$  the production function possesses the following properties:  $\partial Y/\partial K > 0$ ,  $\partial Y/\partial L > 0$ ,  $\partial^2 Y/\partial K^2 < 0$ , and  $\partial^2 Y/\partial L^2 < 0$

### III. Multi-Neutrality and Some Evidence

We have developed a whole series of logically possible and plausible<sup>6</sup> hypotheses about the character of technical progress. Can the available historical record be used to test them? In fact, the regression analyses with linear specifications<sup>7</sup> of single neutrality as performed by Beckmann and Sato on time-series for the United States, Germany, and Japan may be used as evidence about multi-neutralities as developed above. Conclusions derived from the Beckmann-Sato results are summarized in Table 3. What is

answer to these two questions. We also call attention to some apparent implications of the regressions for the good behavior of the production function.

The coefficients  $b_i$  and  $c_i$  of Table 3 refer to parameters in the corresponding linear specifications of single neutrality. For example, the equation to be tested for Hicks neutrality may be written:

$$(9) \quad \frac{y - x \frac{dy}{dx}}{\frac{dy}{dx}} = a_1 + b_1x + c_1t$$

TABLE 3—SUMMARY INTERPRETATION OF REGRESSION RESULTS WITH LINEAR SPECIFICATION OF SINGLE NEUTRALITIES<sup>a</sup>

Type of Neutrality	United States		Coefficients Japan		Germany	
	$b_i$	$c_i$	$b_i$	$c_i$	$b_i$	$c_i$
I	+	$N^{b,c}$	+	$N^{b,c}$	+	$B$
II	+	$N^{b,c}$	+	$B$	+	$N^{b,c}$
III	+	$B$	+	$B$	+	$B$
IV	—	$B$	—	$B$	—	$B$
V	— <sup>b,c</sup>	$N^b$	—	$B$	—	$B$
VI	+	$B$	+	$B$	+	$B$
VII	— <sup>b</sup>	$N^{b,c}$	+	$B$	+	$B$
VIII	— <sup>b,c</sup>	$B$	— <sup>b,c</sup>	$B$	—	$B$
IX	+	$B$	+	$N^{b,c}$	—	$B$

<sup>a</sup> Based on data in Beckmann and Sato (1969) pp. 96-100. Plus and minus signs indicate, respectively, regression coefficients consistent with and inconsistent with anticipated directions of relationships. The letters  $N$  and  $B$  stand, respectively, for neutral and biased.

<sup>b</sup> Regression coefficient not statistically significant at 1 percent level.

<sup>c</sup> Regression coefficient not statistically significant at 5 percent level.

the meaning of these results? What are their implications for multi-neutrality? In the remainder of this section we shall attempt an

<sup>6</sup> They are plausible in the weak sense that if the motives for invention were governed by curiosity alone, one type of neutrality would be as likely to occur as another. Specifically in stating the plausibility of the hypotheses in general we are abstracting from possible systematic influences leading to induced bias in one sense or another.

<sup>7</sup> It would seem that in a number of cases a zero-intercept specification, making the condition even simpler, would be plausible.

If technical progress is, in fact, Hicks-neutral,  $c_1$  should not be significantly different from zero.<sup>8</sup> The letters  $N$  and  $B$  under the columns headed  $c_i$  indicate neutrality or bias under this criterion.

Now in the process of testing for neutrality we also are able to check on the good behavior of the production function. Specification of a well-behaved production function has implications for the direction of the relations between variables specifying single neutralities.<sup>9</sup> Thus if the underlying production function is indeed well behaved, the  $b$  coefficients should have the predicted sign.

In examining the results of the overall

<sup>8</sup> It would also appear, however, that even if the coefficient were significantly different from zero in the statistical sense, from a broader viewpoint and for practical purposes technical progress might be regarded as neutral when the partial elasticity of the dependent variable with respect to the non-time independent variable is large compared with its changes attributable only to time.

<sup>9</sup> The anticipated directions, which may easily be verified graphically, are summarized here for the convenience of the reader:

Pair of Variables	Predicted Direction of Relationship
$r/w, L/K$	+
$r, Y/K$	+
$w, Y/L$	+
$w, Y/K$	—
$r, Y/L$	—
$r/w, Y/K$	+
$w/r, Y/L$	+
$w, L/K$	—
$r, K/L$	—

statistical analysis, what pattern would we expect to see if the "true" production function was well behaved and dual neutral in a compatible sense? First we would expect *all* the  $b$  coefficients to be statistically significant and have the "correct" sign. Secondly we would expect all  $c$  coefficients to be statistically significant (sign being of no special consequence) except for the two contained in the statements of the true single neutralities.<sup>10</sup> Were we to obtain  $c$  coefficients not significantly different from zero in a pair of incompatible specifications for neutral technical progress, it would tend to raise doubts about the reliability of the data, about the specification of the model, or about the statistical estimating procedure.

As may be verified in Table 3, inspection of the  $b$  coefficients reveals that the proposed test tends to cast considerable doubt on the good behavior of the production function. Although there appears some tendency toward consistency with expectations, about 40 percent of the coefficients either are not significantly different from zero or have signs

<sup>10</sup> In some instances all this apparently consistent evidence may after all be involved with logical inconsistency. In particular the dual neutral conditions with linear, zero-intercept specifications may imply a poorly behaved production function. Consider, for example, the production function

$$y = \left( \frac{b}{1-d} \right) x^2$$

which is implied by the conditions

$$y - x \left( \frac{dy}{dx} \right) = g(y)$$

and

$$\frac{dy}{dx} = h(x)$$

with specifications  $g(y) = dy$  and  $h(x) = bx$ . The production function may be written  $Y = (b/(1-b))L^2K^{-1}$  so that

$$\frac{\partial Y}{\partial K} = (-1) \left( \frac{b}{1-b} \right) L^2 K^{-2}$$

and

$$\frac{\partial Y}{\partial L} = (2) \left( \frac{b}{1-b} \right) LK^{-1}$$

Since  $L$  and  $K$  both must be positive it is quite clear that either  $\partial Y/\partial K$  or  $\partial Y/\partial L$  must be negative, contrary to the requirements for good behavior.

different from those implied by good behavior. When we focus on the nontraditional (IV-IX) varieties, the matter appears even more alarming since about 60 percent of the coefficients differ from those anticipated.

Turning attention to compatibility of apparent single neutralities, it seems that the results again are scarcely reassuring. With respect to the United States, the data are consistent not only with Hicks and Harrod neutrality, a compatible pair, but also with Capital-combining and with Capital-decreasing neutralities, an incompatible combination of single types neither of which is consistent with the Hicks or Harrod variety. The data for Japan suggest only one pair of single neutralities, but they too appear on theoretical grounds to be incompatible. Only in the case of Germany do no incompatible pairs appear.

#### IV. Conclusions

In sum we have found first that the apparent proliferation of alternative hypotheses regarding neutrality of technical progress may not lead to such a large and cumbersome array of logical possibilities as might at first be surmised. Only one triple-or-higher neutrality was found to be internally consistent. Of thirty-six possible dual neutralities, we might be inclined to reject twenty-seven on the basis of incompatibility provided, of course, we are willing to retain our faith in the good behavior of the underlying production function. Secondly we can derive a series of general conditions capable of generating families of production functions possessing the property of multi-neutrality in a specified sense. Finally, we have found once again that empirical testing leads to something less than crystal clear grounds for rejecting some plausible logical possibilities. Under the circumstances we hardly can reject with great conviction any of the hypotheses of single neutral or of compatible multi-neutral technical progress.

#### REFERENCES

- R. G. D. Allen, *Macro-Economic Theory*, New York 1968.  
M. J. Beckmann and R. Sato, "Aggregate Pro-

duction Functions and Types of Technical Progress: A Statistical Analysis," *Amer. Econ. Rev.*, Mar. 1969, 59, 88-101.

——— and ———, "Neutral Inventions and Production Functions," *Rev. Econ. Stud.*, Jan. 1968, 35, 57-67.

C. E. Ferguson, *The Neoclassical Theory of Production and Distribution*, Cambridge 1959.

H. Uzawa, "Neutral Inventions and the Stability of Growth Equilibrium," *Rev. Econ. Stud.*, Feb. 1961, 28, 117-24.

# Inflation, Unemployment, and Economic Welfare: Comment

By GORDON TULLOCK\*

Contrary to the orthodox view, Roger Waud argues in his recent article in this *Review* that anticipated inflation "... has an unambiguously positive effect on attempts to achieve full employment ..." (p. 631). His reasoning, however, depends on the use of "a fixed money wage." His arguments for fixity in the money wage (see fn. 3, p. 631) are reasons for believing that wages will not fall in the face of unemployment, not for believing that they will not rise in an anticipated inflation.<sup>1</sup> In fully anticipated inflation, real wages would tend to be fixed, while money wages would rise. Under these circumstances, the argument with respect to unemployment would not follow.<sup>2</sup> This does not affect Waud's argument with respect to the other

two results of inflation. Thus I am objecting to a little less than one-third of his article.

The classical argument for inflation in unemployment situations has assumed that it is not anticipated, and hence that money wages are "sticky" while real wages can be depressed by inflation. Waud strengthens the sticky assumption into one of complete fixity, and hence duplicates the normal argument for unanticipated inflation in an anticipated-inflation context. The point is of considerable importance because it would appear that the long period in which economic policy of an inflationary nature could be used to wipe up unemployment has terminated in England, because inflation is now anticipated. The United States appears to be on the edge of a similar situation, and West Germany is approaching the point of transition. The continuing unemployment in both the United States and England, with considerable rates of inflation, can only be explained on the assumption that real wages are becoming sticky because people are beginning to see through the "money illusion."

## REFERENCES

- M. J. Bailey, "The Welfare Cost of Inflationary Finance," *J. Polit. Econ.*, Apr. 1956, 64, 93-110.  
R. N. Waud, "Inflation, Unemployment, and Economic Welfare," *Amer. Econ. Rev.*, Sept. 1970, 60, 631-41.

\* Center for Study of Public Choice, Virginia Polytechnic Institute and State University.

<sup>1</sup> Martin Bailey specifically assumes that all "... wages ... have regular 'cost-of-living' adjustments or the like" (p. 94). Edmund Phelps and Milton Friedman, the other two authorities cited in Waud's article (fn. 1, p. 631), clearly have the same model, but do not discuss wages explicitly.

<sup>2</sup> Waud returns to the subject of unemployment briefly in the course of discussing other matters (p. 637). There, particularly in his footnote 18, he discusses the possibility that the inflation might not eliminate all employment, and hence that there might be an unemployment equilibrium. This unemployment equilibrium would, however, be lower than the amount of unemployment without the inflation, and the lowering of the unemployment would come because the wage is fixed in nominal terms.

# Inflation, Unemployment, and Economic Welfare: Reply

By ROGER N. WAUD\*

The fixed money wage assumption along with a flexible commodity price is often used to characterize the "complete Keynesian model." This asymmetry means that firms are on their demand schedules for labor, as implied by equation (1) of my article, p. 631, but that labor is not on its supply schedule except when the system is in a full employment position. Alternatively, it could just as well be assumed that both the commodity price and the money wage are inflexible during the period of analysis under consideration; then firms are not on their demand schedules for labor nor are laborers on their supply schedules when the system is at less than full employment positions. This assumption is implicit in all of the pure quantity adjustment type Keynesian models (i.e., where both prices and wages are assumed fixed), and it is therefore implicit that the real wage is assumed fixed. Gordon Tullock suggests that under these circumstances my argument that anticipated inflation keeps the system closer to a full employment position "would not follow."

I will show that under these circumstances the argument still holds, and therefore is not the consequence of the asymmetric assumption about the fixity of the money wage and the perfect flexibility of the price level. This follows from the kind of period analysis which Axel Leijonhufvud (pp. 36-39 and ch. 2) has emphasized in his interpretation of Keynes: "The standard 'Keynes and the Classics' analysis places great stress on the restrictive nature of the 'wage rigidity' assumption. But this strong assumption is not necessary in order to explain system behavior of the Keynesian kind. It is sufficient just to give up the equally strong assumption of instantaneous price adjustments" (p. 37). My contention, p. 631, that anticipated inflation moves the system closer to a full employment position does not depend on the rigid

money wage assumption, as Tullock asserts, but rather on the effects which anticipated inflation has on the demand for real money balances.

Consider once again the model of my article, but now assume that the price level as well as the money wage are inflexible. Using the same notation and equation numbers, the model now consists of equations (4) and (6), pp. 632-33; equation (1) no longer holds since firms are not on their demand schedules for labor. Assuming the anticipated rate of inflation  $\rho$  to be given, the model is

$$(4) \quad m = yk(r + \rho)$$

$$(6) \quad I(r) = S\left(r, \lambda \frac{cy}{r} + m, y, \bar{T}\right) + \bar{T}$$

where the endogenous variables are the real income level  $y$  and the real interest rate  $r$ . Again, the locus of  $y$  and  $r$  which satisfy equation (4) for given  $\rho$ , hence giving equilibrium in the money market, defines the  $MM$  schedule of Figure 1 in my article, p. 633. Similarly, the locus of  $y$  and  $r$  satisfying equation (6), thereby giving equilibrium in the goods market, defines the  $EE$  schedule of Figure 1. As I explained in my article, p. 633, an exogenous increase in the anticipated rate of inflation  $\rho$  will cause the  $MM$  schedule to shift rightwards thereby causing a rise in the level of real output and employment. In this case the real wage is unchanged due to the inflexibility of the money wage and the commodity price.

More generally, the money wage and the price level could be changing at the same percentage rate (equal  $\rho$ ); I refer the reader to footnote 18 of my article, p. 637, as Tullock does in his footnote 2. Again, if firms are not on their demand schedules for labor and laborers are not on their supply schedules, an increase in the anticipated rate of inflation  $\rho$  will lead to a reduction of the unemployment level because the increase in the opportunity cost of holding money causes the

\* University of North Carolina, Chapel Hill.

*MM* schedule to shift rightwards; the real wage could remain unchanged with prices and wages simply increasing at the higher rate of anticipated inflation. Or in Tullock's words, "In fully anticipated inflation, real wages would tend to be fixed, while money wages would rise." It is clear however that it is *not true* that: "Under these circumstances, the argument with respect to unemployment would not follow."

In my article an asymmetry between wage flexibility and price flexibility is assumed, and equation (1) is assumed to hold, i.e., that firms are on their demand schedules for labor while laborers are not on their supply schedules. I grant that the use of this type of model may tend to obscure the fundamental mechanism through which anticipated inflation causes the level of output and employment to be higher; the fixity of the nominal wage is certainly not the *deus ex machina* it might appear to be. If there is one, it is that the real

wage is not at the level, assuming such a level exists, necessary for full employment. As Leijonhufvud convincingly argues, however, the Walrasian auctioneer, implicit behind the assumption of instantaneous wage and price adjustment, is an even more dangerous *deus ex machina* in that the problem of how the system really finds the equilibrium price vector is simply ignored. For whatever reasons the real wage may not adjust to a level consistent with full employment, analyzing the problems resulting from its failure to do so is certainly better than assuming such problems away.

#### REFERENCES

- A. Leijonhufvud, *On Keynesian Economics and the Economics of Keynes*, Oxford 1968.
- R. N. Waud, "Inflation, Unemployment, and Economic Welfare," *Amer. Econ. Rev.*, Sept. 1970, 60, 631-41.

# Lerner on Pollution Controls: Comment

By ALI M. REZA\*

In a recent article concerned mainly with the manner in which an optimum quantity of pollution could be attained, Abba Lerner argues:

There is no justification for compensation in the case of new products, or even new plants in old firms producing old products, so the difficulty [of keeping government subsidies to a minimum] does not exist. Compensation is justified only where a plant was established *before* the pollution control was instituted. The polluter can then claim that the burden of the abatement should fall on the community as a whole that benefits from the abatement and that he should not be liable for more than his share of the taxes to pay the compensation. There is no justification for any claim of compensation for plants set up after the establishment of the pollution control. Those who set them up knew of the pollution charge before they made their investment. [p. 530]

This policy recommendation has serious implications, as the following analysis will demonstrate.

Consider the case where firm  $X$  is contemplating, at time  $t$ , to produce good  $X$ . The firm has knowledge of prices of each of the resources needed for the production of  $X$ ; let us consider one of these prices, namely, the wage rate,  $w(t)$ . The firm expects the wage rate to be  $\hat{w}(x; t+1), \dots, \hat{w}(x; t+n)$  for next  $n$  years. Given these wage rates and actual and expected prices of output and prices of the remaining inputs, the firm proceeds with the production of  $X$ .

In year  $t+k$  ( $k < n$ ), the community in which firm  $X$  operates changes its preferences and decides to surround itself with public parks. The parks, being public goods, will

require unified action. Assuming the quantity of other governmental (that is, community) services to remain constant, and assuming full employment, new taxes are raised and labor is employed for the project (which is expected to take  $m$ ,  $m \geq n-k$ , years to complete). For a less-than-perfectly elastic supply of labor, the increased demand for labor will lead to a higher equilibrium wage rate.

We may surmise that  $k$  years ago, firm  $X$  did not (and probably, could not) have foreseen such a change in tastes; therefore, the firm's estimate of wage rates  $\hat{w}(x; t+k+i)$ ,  $i=1, \dots, n-k$ , would now be considered too low. An upward revision is called for.

Next consider firm  $Y$ , at time  $t+k$ , contemplating the production of good  $Y$ . We do not rule out the possibility of firm  $Y$  being the same firm  $X$  nor the possibility that good  $Y$  might be identical to good  $X$ ; however, it is required that firm  $Y$  build a new plant. Obviously, firm  $Y$  has the new information concerning the higher level of the wage rate (which  $X$  lacked  $k$  years ago). If the same information is available to both firms and if they purchase labor competitively, then the actual wage rate,  $w(t+k)$ , and the expected wage rates  $\hat{w}(t+k+i)$ ,  $i=1, \dots, n-k$ , will be identical for both firms. But we also know that  $\hat{w}(t+k+i) > \hat{w}(x; t+k+i)$  for every future period.

Should the community compensate firm  $X$  by means of a subsidy payment because the firm did not anticipate the increase in the wage rate? And if, for any reason, it was decided to provide firm  $X$  with such a subsidy, why should firm  $Y$  not expect it, too? I do not believe that many economists would suggest that it was justified to pay subsidies to a firm whose costs of production had risen because of a greater demand for products which used the same kind of resources utilized by that firm. It is possible that, due to the unexpected increase in the wage rate, firm  $X$  may have to shut down; but the cause would be

\* Graduate School of Business, University of Pittsburgh. Discussions with Professors M. Spiro and E. Sussna have been of great value. Any errors are, of course, solely my responsibility.

that the price mechanism performed its duties all too efficiently.

The analysis presented above is analogous to the case when a nation becomes involved in a war unexpectedly, which raises cost of production in certain sectors of the economy. And when the conflict is settled, again unexpectedly, the defense-related firms discover that demand for their products has vanished. Is compensation justified on economic efficiency grounds, in either case?

Note that nowhere in my analysis was any specific value assumed for the actual or expected wage rates; the particular values these variables may take are irrelevant for the purpose of this analysis. Therefore, we may assign the value zero to  $w(t)$  and  $w(x; t+k+j)$ ,  $j=1, \dots, n$ , without affecting the analysis or its results. However, this assumption is not necessary for the analysis.

Now substitute the terms air for labor, taxes (or cost of pollution, or damage to the environment) for the wage rate (actual or expected), and clean air for public parks surrounding the community. I do not see anything in the new terminology which leads me to change the analysis or its conclusions.

We have looked at the problem by regarding clean air as a factor of production. We may also view the problem of producing clean air—that is, regard clean air as the output of the production process. Again, nothing changes; this can be demonstrated by viewing reduced labor services to imply an increase in the output of leisure.

Then it may be said that the issue involved is one concerned with income distribution. Profits a firm earns have to do both with the demand and supply schedules faced by the firm. Are we justified to prevent a firm's profits from falling below what it expected them to be because its supply schedule has

shifted (as a result of the institution of pollution controls)? Then, what about the aerospace engineer who is earning far below what he expected years ago when he decided upon his career, simply because tastes have changed? In both instances, anticipations did not realize and both deserve the same type of treatment. Suppose an individual observed the future trend in the aerospace industry and decided to become a computer scientist; perhaps the aerospace engineer would have chosen computer sciences had he been able to postpone his decision of the selection of a career. Are we then justified to redistribute income from one to the other (by taxing one and paying a compensation to the other) simply because one had to make his decisions prior to new developments in the economy?

In conclusion, all polluters should be treated alike. If one tends to regard clean air to be an input to production, all polluters should be charged the cost. On the other hand, if one tends to regard clean air as an output, then its producers should be paid for the good they produce. The former view would consider taxation an appropriate means of pollution control; the latter view would justify the payment of subsidies. In both cases, of course, consumers as a group will bear the costs involved, although the particular scheme adopted will have a bearing on the allocation of resources and incomes within the economy.

#### REFERENCES

- A. P. Lerner, "The 1971 Report of the President's Council of Economic Advisors: Priorities and Efficiency," *Amer. Econ. Rev.*, Sept. 1971, 61, 527-30.
- U.S. Council of Economic Advisors, *Economic Report of the President*, Washington 1971.

# Pollution Abatement Subsidies

By ABBA P. LERNER\*

Ali Reza's note shows that I have not been as careful as one should be when stretching old terminology to fit new uses. A statement that "... There is no justification for any claim of compensation for [new] plants ..." is not intended to mean that there is always or even generally, justification for compensation (in the form of a "pollution abatement subsidy") in the case of old plants. Instead of writing "Compensation is justified only where a plant was established *before* the pollution control was instituted" I should have written "compensation *may be* justified ... etc."

It is true that the institution of pollution control creates unexpected costs, but I would not argue that everybody must be compensated for every unexpected cost that results from every action undertaken by government in the social interest. To do so is to fall into the "new welfare economics" trap which would limit the economists' recommendation to actions in which every hurt is completely compensated. No action (and, indeed, no failure to act) could then be recommended. There is always somebody hurt by the action (or inaction) who cannot be identified and compensated.

I agree that "all polluters should be treated alike" but only in being penalized alike for *marginal* polluting. This condition is satisfied by the appropriate form of "pollution abatement subsidies." I apparently failed to make it clear enough that the appropriate form of such a subsidy is a *lump sum compensation* with a *reduction* of the compensation for each unit of pollutant produced. The reduction of the compensation as more pollution is produced has the same effect as a tax per unit of pollution. A lump sum compensation by itself, like a lump sum tax, has no direct effect on the production of pollutants. It is a pure transfer of income or wealth to compensate one who has shown

that an unfair burden was imposed on him by an action undertaken in the social interest.

Whether compensation should be paid cannot depend on whether one views clean air as an output or as an input. Reducing the output of clean air is equivalent to using up some clean air as an input in producing something else. In both cases, there should be the same penalty, equal to the damage from the resulting decrease in the quantity of clean air. And the same penalty is imposed by a reduction in the amount of subsidy received as by an increase in the amount of tax paid.

The confusion may be the result of treating pollution as if it consisted of *reducing the quantity of a good* by producing less of it. But this is not the nature of pollution. *Pollution is the production of a bad* like sulphur dioxide—not a reduction in the quantity of a good like clean air. In the rare cases where the harm from an activity consists of a reduction in the quantity of some good, we are back with the familiar economic analysis of goods and services. Receiving more "compensation" for using up less clean air is then only an unusual way of saying "being paid for the good," and a tax per unit of clean air used up, is only an unusual way of saying "paying for the good."

Compensation for damage caused by measures imposed for the abatement of pollution may be justified only if the damage is a direct result of a political decision to institute any change in the public interest. Even then it can be justified only as furthering *distributive justice*, as protecting the existing distributional *status quo*, or as *tribute* necessary for obtaining the consent of people who could prevent the change from being made.

Government intervention is called for where a good like clean air is being used up excessively, but this is not because of pollution. No *bad* is being produced. The problem is that the product is a *public good*. Its users use it excessively because they do not have to pay for it. They do not have to pay for it

\* Queens College, City University of New York.

because it is not anybody's property. Everyone waits for others to pay for its production so that it does not get produced unless the government buys it, and nobody tries to economize in its use since he does not have to pay for it unless the government imposes a tax per unit used up. (The tax saving from using up fewer units of clean air is equivalent to being paid for "producing" the unused clean air.)

There is here clearly no reason for discriminating between new and old users (or producers) of clean air. The (per unit) government subsidy on the production, or charges on the use, of a public good is quite different from the (lump sum) compensation part of the pollution abatement subsidy.

Also different is the subsidy required for ideal output in the case where marginal cost is less than average cost. This is a case of negative rent and also applies just as much to new firms or plants as to old ones. This

subsidy is required not on behalf of distributive justice or the status quo or as tribute but only for economic efficiency. Without it, the socially desirable level of production would mean running a loss and so would not take place. (The institution of a pollution tax of so much per unit of pollution raises the marginal cost of the good produced in the course of creating the pollution. It therefore tends to reduce the quantity of the good produced and the *rents* earned in its production. As long as the rents remain positive any pollution abatement subsidy (lump sum subsidy accompanying the tax per unit of pollution created) must be for the sake of justice, the status quo, or tribute. It is, however, possible that the pollution tax transforms the positive rents into negative rents. In that case, ideal output may call for a further government subsidy to absorb the negative rent.)

# Price-Quantity Adjustments in a Competitive Market

By E. C. H. VEENDORP\*

Disequilibrium in a competitive market obviously implies that one or more traders are unable to buy or sell as many units of a commodity as they would like to at the current market price. To prevent frustrations on the part of these individuals, it is usually assumed that they can recontract. Unfortunately, this assumption obscures most of the interesting implications of disequilibrium situations.

Several authors have recently analyzed the adjustments in supply and demand made by individual traders in response to disequilibrium situations.<sup>1</sup> Conducted within the context of general equilibrium models, these studies concentrate on the crucial but intricate phenomenon of spillover by which the nonrealization in one market may affect the excess demands in all other markets. While considerable progress has been made along these lines, many questions remain unanswered.

Much less attention has been devoted to adjustments toward equilibrium in one, isolated market. Use of such a simplified framework admittedly excludes the interesting interactions between different markets. It leaves, however, some scope for disequilibrium adjustments in demand and supply, the analysis of which may increase our understanding of more general cases.

## I

A detailed analysis of price-quantity adjustments and their convergence to equilibrium for a single market has been given by Martin Beckmann and Harl Ryder. Their basic model is presented as a combination of a Walrasian price mechanism

$$(1) \quad DP = \lambda(Q_d - Q) \quad (\lambda > 0)$$

\* Associate professor of economics, Tulane University. I wish to thank J. R. Moroney and a referee for helpful comments on an earlier draft of this paper.

<sup>1</sup> See Robert Clower, Axel Leijonhufvud, Herschel Grossman, and Robert Barro and Grossman.

and a Marshallian quantity adjustment process

$$(2) \quad DQ = \mu(P - P_s) \quad (\mu > 0)$$

where  $P$  denotes price,  $Q_d = Q_d(P)$  quantity demanded,  $Q$  quantity supplied, and  $P_s = P_s(Q)$  marginal cost for the industry (supply price). The  $D$  operator indicates differentiation with respect to time:  $D \equiv d/dt$ . The first equation specifies price adjustments proportional to excess market demand. Demand is assumed to adjust instantaneously to this market price, whereas sellers adjust the quantity supplied in proportion to the difference between price and marginal cost (equation (2)).

The second equation is perfectly plausible within its original Marshallian framework, where price adjustments are assumed to be instantaneous so that market demand and supply are continuously equal. Used in conjunction with a Walrasian price mechanism, however, the interpretation of equation (2) is much less satisfactory. Markets no longer clear continuously, and market supply may exceed demand in which case sellers are unable to sell the desired amount. But these sellers will, according to equation (2), continue to increase their supply as long as price exceeds marginal cost, no matter what percentage of their supply can be sold.

It is clear, therefore, that the Beckmann-Ryder model neglects the implications of the possible nonrealization of supply. This neglect would seem to be particularly serious for perishable, and other nonstorable, commodities. The modification proposed in the next section is meant to remedy this shortcoming.

## II

Consider a purely competitive market for a perishable commodity. We assume that the individual buyer or seller accepts the price as a microdatum: he acts as if he had no influence on this price, and merely adjusts

to a given market situation.<sup>2</sup> Whenever price and output levels are such that market supply exceeds demand ( $Q > Q_d$ ), a principle is needed by which demand is allocated among individual sellers. Let us first assume that the market is organized in such a way that all suppliers can sell the same fraction,  $r (= Q_d/Q)$ , of their market supply. It seems plausible that sellers in a purely competitive market organized along these lines will consider this realization rate as another microdatum.<sup>3</sup> The fact that only 100  $r$  percent of their supply can be sold, whereas the remainder disappears by spoilage, implies that for every unit produced they receive only a fraction  $r$  of the (disequilibrium) market price. One would expect sellers to increase their supply if and only if this fraction exceeds marginal cost. We propose, therefore, to replace (2) by

$$(3a) \quad DQ = \mu(P - P_s) \quad \text{if } Q_d \geq Q$$

$$(3b) \quad DQ = \mu(rP - P_s) \\ = \mu(PQ_d/Q - P_s) \quad \text{if } Q_d < Q$$

The same reformulation is appropriate for a market where participants are being treated on a "first come, first served" basis. Assume that all sellers produce the same fraction of total market supply. If supply exceeds demand, 100  $r$  percent of them will succeed in selling their supply at the prevailing market price, and can be counted on to adjust their supply on the basis of the difference between marginal revenue ( $P$ ) and cost. Latecomers (100  $(1-r)$  percent of the number of suppliers) will be unable to sell any units, will experience a zero marginal revenue, and will adjust their production accordingly. Taking the weighted average of these adjustments, we have, for excess supply situations,

$$DQ = r\mu(P - P_s) + (1-r)\mu(0 - P_s) \\ = \mu(rP - P_s)$$

which is the same as equation (3b).

<sup>2</sup> The use of such an impersonal, auctioneer-type market mechanism has correctly been criticized by Kenneth Arrow and others. This is a separate issue, however, that will not be considered here.

<sup>3</sup> For the use of realization rates in general equilibrium theory, see Veendorp, pp. 13-20.

### III

To analyze the effects of the modification proposed in the preceding section on the dynamic properties of the adjustment process, we postulate linear demand and cost curves with "normal" slopes

$$(4) \quad Q_d = a - bP \quad (a, b > 0)$$

$$(5) \quad P_s = c + dQ \quad (d > 0)$$

Equilibrium prevails when demand equals supply ( $Q_d = Q$ ) and price equals marginal cost ( $P = P_s$ ), which happens at

$$\bar{P} = (c + ad)/(1 + bd)$$

$$\bar{Q} = (a - bc)/(1 + bd)$$

Substituting (4) and (5) into equations (1) and (2), the Beckmann-Ryder system reduces to

$$(6) \quad DP = \lambda(a - bP - Q)$$

$$(7) \quad DQ = \mu(P - c - dQ)$$

The phase diagram for this system of differential equations is given in Figure 1. The locus  $DP = 0$  represents price-quantity combinations for which excess market demand is zero; the locus  $DQ = 0$  those for which price equals marginal cost. Due to the nature of the adjustment process these loci have the same shape as the demand and supply curves given by equations (4) and (5). Together they divide the relevant quadrant into four regions, labeled I through IV.

The arrows in the phase diagram indicate the direction of the dynamic forces operating on price and output levels. These directions depend on the signs of the partial derivatives of the two loci with respect to  $P$  and  $Q$ , and assure that the Beckmann-Ryder system is stable for all positive values of  $b$  and  $d$ , and the adjustment coefficients.<sup>4</sup>

<sup>4</sup> See, for example, James Quirk and Richard Ruppert. Figure 1 represents one of the "sign stable" cases in the classification given by Quirk and Ruppert, who apply a theorem of Czeslaw Olech to the general class of qualitatively specified systems of two differential equations. The present system being linear, its stability can also be concluded from the fact that its roots have negative real parts. These roots,  $X$ , are determined by the characteristic equation

$$X^2 + (\lambda b + \mu d)X + \lambda\mu(1 + bd) = 0$$

and have negative real parts for all positive values of  $b$ ,  $d$ ,  $\lambda$ , and  $\mu$ .

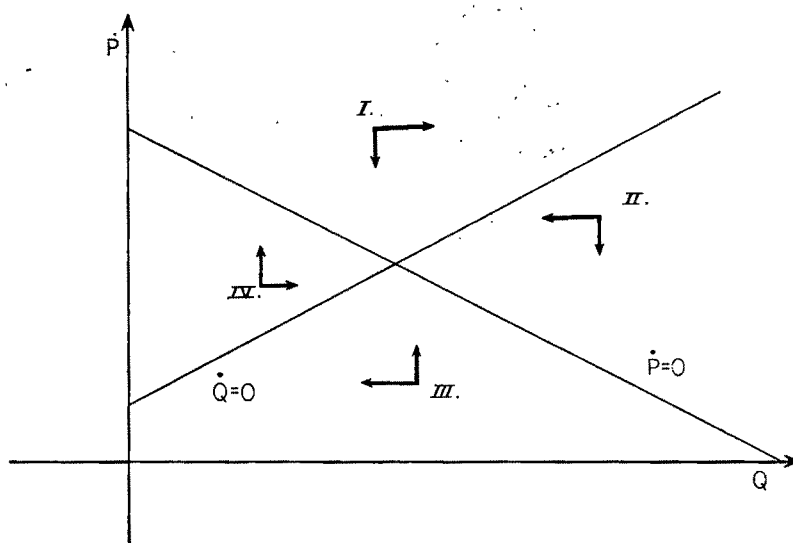


FIGURE 1. EQUATIONS 1 AND 2

## IV

The modification proposed in Section II is effective only in excess supply situations. If supply exceeds demand, the price-quantity adjustments in our system are specified by equations (1) and (3b). Substitution of (4) and (5) reduces this system of differential equations to

$$\begin{aligned} DP &= \lambda(a - bP - Q) \\ DQ &= \mu(rP - c - dQ) \\ &= \mu\{P(a - bP)/Q - c - dQ\} \end{aligned}$$

Unlike equations (6) and (7), this system is non-linear.

To construct the corresponding phase diagram we first rewrite the locus  $DQ=0$  as

$$P(a - bP) - Q(c + dQ) = 0$$

or

$$\begin{aligned} b(P - a/2b)^2 + d(Q + c/2d)^2 \\ = (a^2/b + c^2/d)/4 \end{aligned}$$

This is the equation of an ellipse with center  $P^0 = a/2b$ ,  $Q^0 = -c/2d$ , which goes through the origin and the equilibrium point (Figure 2). It intersects the zero excess demand locus ( $DP=0$ ) at  $Q=0$ , and the linear segment of the locus  $DQ=0$  (or its extension in the

second quadrant) at  $P=0$ . The regions defined by the relevant parts of these loci are numbered as before.

The slope of the ellipse at the equilibrium point may be negative or positive. It is negative if  $\bar{P} > P^0$  (or  $(c + ad)/(1 + bd) > a/2b$ ), and positive otherwise. To simplify the argument we shall assume this slope to be negative, which is the case depicted in Figure 2.

Inspection of the phase diagram in Figure 2 shows that the corresponding adjustment process is stable in the large. Any trajectory starting in regions II, III or IV will move into region I within a finite amount of time or will approach equilibrium without reaching the boundaries of region I. Having moved into region I, any path will remain in that region since the zero excess demand locus serves as a lower bound for price variations, whereas the relevant part of the ellipse serves as an upper bound for quantity variations. The forces operating on price and quantity in region I assure that any such trajectory will approach equilibrium.<sup>5</sup>

<sup>5</sup> To prove stability is less straightforward if the equilibrium point lies on the positively sloped segment of the ellipse. Such a positively sloped segment does not serve as an upper bound for quantity variation, so that a trajectory may leave region I by crossing that segment of its boundary with region II. Being that close to equi-

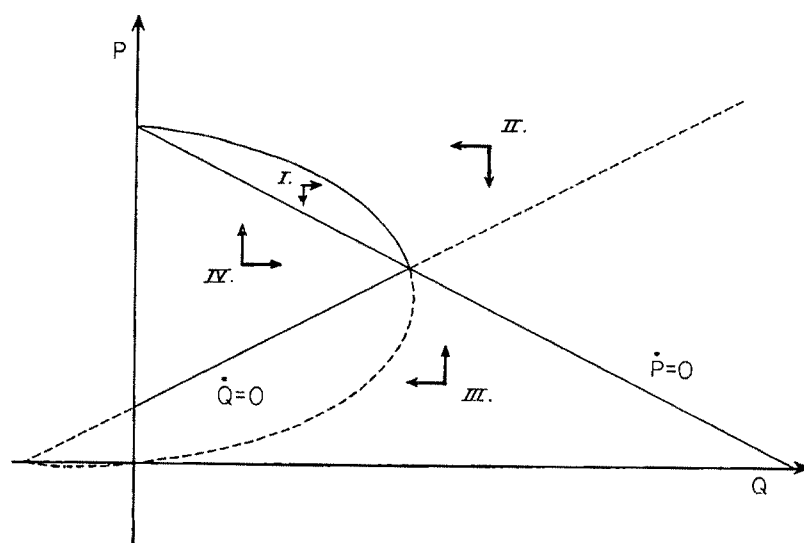


FIGURE 2. EQUATIONS 1 AND 3

## V

Comparing the phase diagrams in Figures 1 and 2 we observe that the modification proposed in this paper affects the relative size of regions I and II. The reduced size of region I reflects the fact that the nonrealization of supply weakens the incentive for sellers to increase output levels. Sellers are still tempted to increase supply when price exceeds marginal cost, but they will resist this temptation if realization rates are too low.

As we have shown in the preceding section, our modified adjustment process preserves the stability of the linear version of the Beckmann-Ryder model.<sup>6</sup> But the approach toward equilibrium may be quite different for the two systems. The modification is bound to affect all trajectories that generate excess supplies during certain periods of time. Its effect is most pronounced on those

librium, however, the dynamic forces operating on price and quantity adjustments are such that stability would seem to follow from Olech's theorem and the analysis of Quirk and Ruppert (see fn. 5). The theorem of Olech assumes that the two loci ( $\dot{D}P=0$  and  $\dot{D}Q=0$ ) have continuous first partial derivatives with respect to the two variables. The fact that this condition is not met at equilibrium would seem to be of no practical significance.

<sup>6</sup> There is no assurance that such equivalence would extend to the more general cases considered by Beckmann and Ryder.

solutions of the Beckmann-Ryder system that represent oscillations around equilibrium.

To illustrate this effect, let the roots of the Beckmann-Ryder system be conjugate complex.<sup>7</sup> Such roots assure a converging clockwise movement around equilibrium in Figure 1. Price and quantity will oscillate around their equilibrium value and periods of excess supply and demand will alternate in a regular fashion. Under the modified conditions of Figure 2 any such clockwise movement comes to an end in region I. Having entered this region the adjustment path starts to approach equilibrium from above and from the left. During this second stage of the adjustment process, price will continuously exceed marginal cost, and market supply will continuously exceed demand, but the excess supplies are insufficient to keep sellers from increasing their output levels. The smaller region I is (or the steeper the supply curve, and the less steep the demand curve), the more closely will this movement correspond to convergence along the demand curve.<sup>8</sup>

<sup>7</sup> This requires  $(\lambda b - \mu d)^2 - 4\lambda\mu < 0$ . See fn. 4.

<sup>8</sup> If the equilibrium point lies on the positively sloped segment of the ellipse, this movement through region I is likely to be succeeded by a third stage during which the variables may start to oscillate again around their equilibrium values. See fn. 5.

REFERENCES

- K. J. Arrow, "Toward a Theory of Price Adjustment," in M. Abramovitz, ed., *The Allocation of Economic Resources*, Stanford 1959.
- R. J. Barro and H. I. Grossman, "A General Disequilibrium Model of Income and Employment," *Amer. Econ. Rev.*, Mar. 1971, 61, 82-93.
- M. J. Beckmann and H. E. Ryder, "Simultaneous Price and Quantity Adjustment in a Single Market," *Econometrica*, July 1969, 37, 470-84.
- R. Clower, "The Keynesian Counter-Revolution: A Theoretical Appraisal," in F. H. Hahn and F. P. R. Brechling, eds., *The Theory of Interest Rates*, London 1965.
- H. I. Grossman, "Theories of Markets Without Recontracting," *J. Econ. Theor.*, Dec. 1969, 1, 476-79.
- A. Leijonhufvud, *On Keynesian Economics and the Economics of Keynes*, New York 1968.
- C. Olech, "On the Global Stability of an Autonomous System on the Plane," *J. Differential Equations*, 1963, 1, 389-400.
- J. P. Quirk and R. W. Ruppert, "Phase Diagrams and Global Stability," Research Papers in Theoretical and Applied Economics, paper #11, Univ. Kansas.
- E. C. H. Veendorp, "General Equilibrium Theory for a Barter Economy," *Western Econ. J.*, Mar. 1970, 1970, 8, 1-23.

# NOTES

At the invitation of the Chinese Academy of Sciences, the Chinese Scientific and Technical Association and the Department of Economics of the University of Peking, the three most recent presidents of the American Economic Association, Professors Wassily Leontief, James Tobin, and John Kenneth Galbraith, visited China in September 1972. The visit included seminars on the Chinese and American economies and an intensive round of visits to commercial and industrial establishments, communes, hospitals, universities, and schools. The Chinese hospitality was warm, the meetings informative and all arrangements showed great thought and attention. The three American participants all expressed their hope that the trip was a portent of further and expanded scientific and cultural exchange.

A Summer Institute for college teachers of economics will be held at Brown University from June 18 to July 27, 1973. The institute is supported by a grant from the National Science Foundation. The program is designed to acquaint teachers in undergraduate institutions with recent developments and policy applications of micro-economic theory. For application forms and further information, write Professor Mark B. Schupack, NSF Institute in Economics, Department of Economics, Brown University, Providence, Rhode Island 02912.

The International Economic Association, with the support of the Ford Foundation, is organizing a Workshop in Mathematical Economics, to be held from June 17 to July 14, 1973 in Schloss Rheda (University of Bielefeld), West Germany. Under the direction of Professor C. C. von Weizsäcker, the workshop will be exclusively in English for a maximum of 25 European scholars and lecturers, preferably under 30 years of age. Applications, with curriculum vitae, academic training and background in spoken English and mathematics, and a note on recent research activities of the applicant, should be addressed to Professor C. C. von Weizsäcker, Institut für Mathematische Wirtschaftsforschung, Universität Bielefeld, 4840 Rheda-Schloss, Germany. Information may also be obtained from the IEA Secretariat, 54 Boulevard Raspail, Paris 6ème. Applications should if possible be submitted before the end of 1972.

## *Call for Papers*

The second annual Intersociety Transportation Conference will be held September 24-27, 1973 at the Brown Palace Hotel, Denver, Colorado. Technical sessions are being planned to cover all aspects of transportation. Prospective authors should submit a 500-word summary (by December 15, 1972, if possible) to the ICT Technical Program Chairman, Mr. Thomas P. Wall, U.S. Department of Transportation, 400-7th Street S.W.,

Room 5408, Washington, D.C. 20590. Paper deadline is April 1, 1973.

Papers are solicited for the fifth Amos Tuck School Seminar on Problems of Regulation and Public Utilities, sponsored by the American Telephone and Telegraph Company, to be held in late August 1973 at Dartmouth College. The seminar provides a forum for the presentation and discussion of new ideas and innovative work in analyzing regulated firms, regulatory processes, and problems of public welfare. We invite those with an interest in this area to suggest names of possible speakers or to apply themselves. The aim of the seminar is to introduce the research both of young scholars beginning their careers and of scholars who have only recently turned their attention to public utility problems. Authors of papers will receive up to \$1,200 in support of research and writing, in addition to travel and living expenses at the seminar. Please send suggestions, biographical information, and samples of written work, including a proposal to be developed for presentation to the directors of the seminar before February 23, 1973.

This is not a request for conferee applications. Application forms and further information will be mailed to institutions in March 1973. The directors of the seminar are: Professors Richard S. Bower and Willard T. Carleton, The Amos Tuck School of Business Administration, Dartmouth College, Hanover, New Hampshire 03755.

On Thursday, May 17, 1973, the Economics Department of the City College of the City University of New York will sponsor an all-day conference on "The Costs and/or Benefits of Education." The object of the conference is to stimulate research that is relevant to current educational issues and policies. Persons interested in reading papers at the conference are invited to submit them to Professor Robert D. Leiter, Department of Economics, City College of the City University of New York, Convent Avenue and West 138th Street, New York, N. Y. 10031. Papers delivered at the conference are scheduled for publication in the first annual conference volume of the department.

## *An Invitation to Join the Society of Government Economists*

As government's role in economic activities becomes more important, so does the economist's role in government. Those individuals who are interested in promoting the scientific application of economics in governmental decision making are invited to write for information about the new Society of Government Economists. Membership is international with emphasis on those who are employed or identified with a governmental

unit. A subscription fee of \$5 (\$10 fee includes a subscription to the *Occasional Papers* and a copy of the membership directory) may be sent or membership forms obtained by contacting Georgia Canellos, Executive Secretary, Society of Government Economists, 400 Seventh Street, S.W., Room 10232, Washington, D.C. 20590. Telephone 202-426-4420.

SGE will hold a symposium in Toronto, December 26-27, 1972 as part of the annual meeting of the Allied Social Science Associations. A conference is also held in Washington each year. Copies of proceedings from previous symposia and conferences can be ordered by completing a publications order form obtained from the Executive Secretary. Economists who wish to submit papers for publication or presentation are invited to submit abstracts or completed papers to the above address for consideration by the appropriate committee chairman.

The Southern Regional Science Association announces election of officers for 1972-73. President: William A. Schaffer, Georgia Institute of Technology; president-elect: Alan R. Winger, University of Florida; secretary-treasurer, James C. Hite, Clemson University. The Association's annual meeting will be held in New Orleans, Louisiana, April 26-27, 1973. Persons interested in participating are invited to contact the program chairman: Professor Alan R. Winger, Department of Economics, University of Florida, Gainesville, Florida 32601.

The Edward M. Ryan prize of \$10,000 has been established to encourage and promote applied research to aid in the solution of existing problems of malnutrition in low income countries. Efforts to combat nutritional problems effectively and on a large scale, however, have been hampered by the lack of adequate information on alternative approaches, suggesting the need for additional operationally oriented research to fill these gaps. For information, write to Dr. F. James Levinson, Chairman, The Edward M. Ryan Prize Committee, The James and Rachel Levinson Foundation, P.O. Box 1617, Pittsburgh, Pennsylvania 15230.

The Joint Committee on *Eastern Europe* of the American Council of Learned Societies and the Social Science Research Council wishes to draw attention to three of its grant programs: grants for postdoctoral research to a maximum of \$8500; grants for study of East European languages to a maximum of \$1000; grants ranging between \$2000 and \$5000 for the support of conferences. For details of eligibility and information which must be supplied in requesting application forms, write to Office of Fellowships and Grants, American Council of Learned Societies, 345 East 46 Street, New York, New York 10017.

The National Tax Association announces the 1972 award winners in the annual competition for outstand-

ing doctoral dissertations in government finance and taxation. The \$1,500 first prize award was won by Daniel L. Rubinfeld of MIT, with his entry, "An Econometric Analysis of Credit Ratings and the Market for General Obligation Municipal Bonds." Honorable mention awards of \$500 each were won by Gerald E. Auten of University of Michigan, "The Measurement of Local Public Expenditure Needs," and George E. Garrison of Clark University, "Tax-Expenditure Analysis of Public Education." The members of the 1972 Selection Committee were Professors E. Cary Brown, Arthur D. Lynn, Jr., James A. Papke, and H. Clyde Reeves. Information on the 1973 award competition may be obtained from Professor James A. Papke, Department of Economics, Krannert Graduate School of Industrial Administration, Purdue University, West Lafayette, Indiana 47906.

### Deaths

Clarence E. Ayres, professor emeritus, University of Texas at Austin, July 24, 1972.

C. Harry Kahn, professor of economics, Rutgers College, Apr. 28, 1972.

Kenneth K. Kurihara, distinguished professor of economics, State University of New York at Binghamton, June 12, 1972.

Maxwell Obst, England, June 18, 1972.

### Retirements

Richard B. Heflebower, professor of economics, Northwestern University, July 1972.

Helen H. Lamale, chief, division of living conditions studies, Bureau of Labor Statistics, U.S. Department of Labor, June 30, 1972.

James G. Maddox, professor emeritus, North Carolina State University, Feb. 1, 1972.

Grover A. J. Noetzel, professor and former dean, department of economics, University of Miami, May 31, 1972.

Alden J. Plumley, professor emeritus, economics department, College of Business Administration, University of Nevada, July 1, 1972.

Ronald B. Shuman, professor emeritus and curator emeritus of the Bass Collection of Business History, University of Oklahoma, June 1972.

T. Wilmot Wood, professor emeritus, North Carolina State University, July 1, 1972.

### Visiting Foreign Scholars

P. N. Junankar, University of Essex: visiting assistant professor, Northwestern University.

Søren Lemche: visiting assistant professor, Northwestern University.

Shlomo Maital, Tel Aviv University: visiting assistant professor, Queen's University, 1972-73.

Ivor F. Pearce, University of Southampton, England: visiting professor of economics, University of California, San Diego, winter and spring quarters, 1973.

*Promotions*

M. Mukhtar Ali: associate professor of economics, University of Kentucky, July 1, 1972.

David L. Anderson: assistant professor, department of economics, Northeastern University, July 1, 1972.

Gerald D. Bell: professor of business administration, University of North Carolina, Chapel Hill, July 1, 1972.

Mario F. Bognanno: associate professor, Industrial Relations Center, University of Minnesota.

Bernard H. Booms: associate professor, department of economics, Pennsylvania State University, July 1, 1972.

Phillips D. Brooks: associate professor of economics, College of Business Administration, University of Nevada, July 1, 1972.

R. Charles Brooks: extension professor of economics, North Carolina State University, July 1, 1972.

William E. Cage: associate professor of economics, Wake Forest University, Sept. 1, 1972.

Charlotte A. Chamberlain: assistant professor, department of economics, Northeastern University, July 1, 1972.

Albin J. Dahl: professor in economics, College of Business Administration, University of Nevada, July 1, 1972.

Charles D. DeLorme: associate professor of economics, University of Georgia.

Alan J. Donziger: assistant professor of economics, Villanova University, fall 1972.

Ronald G. Ehrenberg: associate professor of economics, University of Massachusetts, Sept. 1972.

Nabil A. El-Ramly: associate professor of business economics and quantitative methods, University of Hawaii.

Thomas M. Freeman: associate professor and associate director, Office of Institutional Research, Michigan State University, July 1, 1972.

Klaus Friedrich: associate professor, department of economics, Pennsylvania State University, July 1, 1972.

Alexander Garcia: professor of economics and finance, Fairleigh Dickinson University, Sept. 1, 1972.

Bruce L. Gardner: associate professor of economics, North Carolina State University, July 1, 1972.

Shaikh M. Ghazanfar: associate professor of economics, University of Idaho, fall 1972.

Michael Gordon: assistant professor of economics, State University of New York at Canton, Sept. 1, 1972.

Clifton M. Grubbs: professor of economics, University of Texas at Austin.

Dewey G. Harwood, Jr.: extension professor of economics, North Carolina State University, July 1, 1972.

John D. Hogan: adjunct professor of economics, Ohio State University.

Teh-wei Hu: professor, department of economics, Pennsylvania State University, July 1, 1972.

John W. Isbister: associate professor of economics, Merrill College, University of California, Santa Cruz, July 1, 1972.

James E. Jonish: associate professor of business economics and quantitative methods, University of Hawaii.

Hyman Joseph: associate professor, department of

economics, University of Iowa, Sept. 1, 1972.

Marvin G. Julius: professor of economics, Iowa State University.

David E. Kaun: professor of economics, Stevenson College, University of California, Santa Cruz, July 1, 1972.

Robert B. McBurney: assistant professor of economics, North Carolina State University, July 1, 1972.

Curtis P. McLaughlin: professor of business administration, University of North Carolina, Chapel Hill, July 1, 1972.

Charles E. McLure, Jr.: professor of economics, Rice University, July 1972.

Gene A. Mathia: professor of economics, North Carolina State University, July 1, 1972.

G. Richard Meadows: assistant professor, department of economics, University of Wisconsin-Milwaukee, Sept. 1972.

George T. Milkovich: associate professor, Industrial Relations Center, University of Minnesota.

Richard A. Miller: professor of economics, Wesleyan University, July 1, 1972.

Basil J. Moore: professor of economics, Wesleyan University, July 1, 1972.

Sarah S. Montgomery: professor of economics, Mount Holyoke College.

James L. Murphy: professor of economics, University of North Carolina, Chapel Hill, July 1, 1972.

Henry Orion: associate professor of economics and finance, Fairleigh Dickinson University, Sept. 1, 1972.

Richard K. Perrin: associate professor of economics, North Carolina State University, July 1, 1972.

Richard E. Peterson: associate professor of business economics and quantitative methods, University of Hawaii.

Nallapu N. Reddy: associate professor of economics, Clarkson College of Technology, July 1, 1972.

Clark W. Reynolds: professor, Food Research Institute, Stanford University.

Gaston V. Rimlinger: professor of economics, Rice University.

James Rodgers: associate professor, department of economics, Pennsylvania State University, July 1, 1972.

Owen Sauerlender: professor, department of economics, Pennsylvania State University, July 1, 1972.

Hirofumi Shibata: professor of economics, University of Kentucky, July 1, 1972.

Andrew M. Sum: assistant professor, department of economics, Northeastern University, July 1, 1972.

Vincent J. Tarascio: professor of economics, University of North Carolina, Chapel Hill, July 1, 1972.

Robert W. Thomas, Jr.: professor of economics, Iowa State University.

Rinaldo Toporovsky: assistant professor of economics and finance, Fairleigh Dickinson University, Sept. 1, 1972.

Carl B. Turner: professor of economics, North Carolina State University, July 1, 1972.

Roger N. Waud: professor of economics, University of North Carolina, Chapel Hill, July 1, 1972.

Raburn M. Williams: associate professor of business economics and quantitative methods, University of Hawaii.

Jan P. Wogart: associate professor, department of economics, University of Miami, June 1, 1972.

Pan A. Yotopoulos: professor, Food Research Institute, Stanford University.

Mahmood A. Zaidi: professor, Industrial Relations Center, University of Minnesota.

### *Administrative Appointments*

Robert F. Adams: assistant chancellor, University of California, Santa Cruz.

Virete Angkatavanich: chairman, department of economics and finance, Fairleigh Dickinson University, Sept. 1, 1972.

Richard E. Attiyeh: chairman, department of economics, University of California, San Diego, July 1, 1972.

William M. Baird: associate dean, College of Wooster, June 1972.

Jerald R. Barnard: chairman, department of economics, University of Iowa, July 1, 1972.

Jacob Ben-Moshe: deputy chairman, department of economics and finance, Fairleigh Dickinson University, Sept. 1, 1972.

Philip M. Carroll: director, graduate business programs, College of Administrative Science, Ohio State University, July 1972.

Joseph S. DeSalvo: chairman, department of economics, University of Wisconsin-Milwaukee, Aug. 28, 1972.

Walter P. Falcon: Harvard University Development Advisory Service: director and professor, Food Research Institute, Stanford University.

Samuel Gubins: chairman, department of economics, Haverford College, June 1972.

J. William Hanlon, Winona State College: associate professor of economics and executive director, Georgia Council on Economic Education, Georgia State University, Sept. 1, 1972.

John A. Haslem: chairman, division of finance, department of business administration, University of Maryland, July 1972.

Yaqub N. Karkar: chairman, department of economics and business administration, University of Wisconsin Center System, 1972-73.

Abdul G. Khan: head, programme on institutional management in higher education, Center for Educational Research and Innovation, OECD, Paris.

Melvin Lurie: associate dean, College of Letters and Science, University of Wisconsin-Milwaukee, July 1, 1972.

Gary A. Lynch: chairman, department of economics, University of Idaho, fall 1972.

Stephen L. McDonald: chairman, department of economics, University of Texas at Austin.

Saul Mason: chairman, department of economics, Villanova University, Sept. 1, 1972.

M. David Merchant: project director, International and Governmental Affairs, American Association of Collegiate Schools of Business, July 1, 1972.

Hugh O. Nourse: chairman, department of economics, University of Missouri, St. Louis, Aug. 30, 1972.

Robert H. Persons, Jr.: chairman, department of

economics, University of Bridgeport, July 1, 1972.

Richard Reimer: chairman, department of economics, College of Wooster, Sept. 1972.

Gaston V. Rimlinger: professor of economics and chairman of the department, Rice University.

Robert L. Robertson: chairman, department of economics, Mount Holyoke College, July 1, 1972.

Delwin A. Roy, Ford Foundation: chairman, graduate program in development administration, American University, Beirut, July 1, 1972.

Daniel B. Suits: deputy provost, Merrill College, University of California, Santa Cruz.

### *Appointments*

Sheila A. Adams: research analyst, bureau of business and economic research, College of Business Administration, University of Nevada, July 1, 1972.

Irma Adelman: professor of economics, University of Maryland, Sept. 1972.

Syed Ahmad, University of Kent: professor of economics, McMaster University.

William H. Albright, USAF: staff member, management sciences department, The Rand Corporation, Mar. 1972.

Steven Andersen: lecturer in economics, Rutgers College, 1972-73.

Thomas C. Anderson: assistant professor, department of economics, Eastern Michigan University, June 1972.

Terry D. Aranoff: lecturer in economics, Rutgers College, 1971-73.

Swarnjit Arora, State University of New York at Buffalo: assistant professor, department of economics, University of Wisconsin-Milwaukee, Sept. 1972.

Calvin W. Atwood: assistant director, MBA program, University of North Carolina, Chapel Hill, June 1, 1972.

William E. Avera: assistant professor of business administration, University of North Carolina, Chapel Hill, August 1, 1972.

Joseph Baldwin: visiting professor of economics, University of North Carolina, Chapel Hill, 1972-73.

William R. Barnes: instructor of economics, North Carolina State University, Aug. 1972.

A. P. Baroutsis, Purdue University: associate professor, department of business and economics, Slippery Rock State College, June 1972.

John D. Bazley: assistant professor of accounting, University of North Carolina, Chapel Hill, Aug. 1, 1972.

Rex O. Bennett: visiting lecturer in business administration, University of North Carolina, Chapel Hill, 1972-73.

Miles O. Bidwell, Columbia University: assistant professor of economics, Wake Forest University, Sept. 1, 1972.

William C. Birdsall: visiting associate professor of economics, McMaster University.

Stanley W. Black: associate professor of economics, Vanderbilt University, Sept. 2, 1972.

P. A. Boyer, Old Dominion University: associate professor, department of business and economics, Slippery Rock State College, Sept. 1972.

James J. Clarke: assistant professor of economics, Villanova University, fall 1972.

Julie S. DaVanzo: University of California, Los Angeles: staff member, economics department, The Rand Corporation, Apr. 1972.

Flora Davidov, Stanford University: acting assistant professor of economics, Cowell College, University of California, Santa Cruz, July 1, 1972.

Allan C. DeSerpa: assistant professor of economics, Louisiana State University, Aug. 24, 1972.

Thomas P. Drinka: research associate, department of economics, Iowa State University.

Robert V. Eagly: visiting professor of economics, University of North Carolina, Chapel Hill, 1972-73.

Dawn E. Elvis: visiting assistant professor of economics, Vanderbilt University, Sept. 1, 1972.

Thomas J. Espenshade, University of California, Berkeley: assistant professor of economics, Bowdoin College, Sept. 1, 1972.

Craig B. Foch, Johns Hopkins University: staff member, economics department, The Rand Corporation, Sept. 1972.

Peter Fortune: lecturer, department of economics, Northeastern University, Sept. 1972.

Carolyn A. Fost, Kentucky Wesleyan College: assistant professor, department of economics, Western Kentucky University, Aug. 1972.

Donald E. Frey, Princeton University: assistant professor of economics, Wake Forest University, Sept. 1, 1972.

H. Frederick Gallasch, Jr.: instructor of economics, North Carolina State University, Aug. 1972.

Peter T. Gottschalk, Williams College: assistant professor of economics, Stevenson College, University of California, Santa Cruz, July 1, 1972.

Peter J. Grandstaff: assistant professor of economics, University of Missouri-St. Louis, Aug. 30, 1972.

Paul R. Gregory: associate professor, department of economics, University of Houston, Sept. 1972.

Georg Hasenkamp, University of Wisconsin: assistant professor, department of economics, Wayne State University, Sept. 1972.

Daryl A. Hellman: assistant professor, department of economics, Northeastern University, Sept. 1972.

George D. Hughes, Burlington Industries: professor of business administration, University of North Carolina, Chapel Hill, July 1, 1972.

Thomas R. Ireland: associate professor, department of economics, University of Missouri-St. Louis, Aug. 30, 1972.

George D. Irwin: USDA professor, North Carolina State University, July 1, 1972.

Carl P. Jordan, Columbia University: assistant professor of economics, Frostburg State College, Sept. 1972.

S. Daniel Kanyr: instructor of economics, North Carolina State University, Aug. 1972.

Donald B. Keesing: professor of economics, University of North Carolina, Chapel Hill, Aug. 1, 1972.

David L. Kelly: visiting lecturer in business administration, University of North Carolina, Chapel Hill, 1972-73.

J. Thom Kelly: post-doctoral fellow in economics, University of North Carolina, Chapel Hill, 1972-73.

Mark R. Killingsworth: visiting assistant professor of economics, Vanderbilt University, Sept. 1, 1972.

Sungwoo Kim: associate professor, department of economics, Northeastern University, Sept. 1972.

Jan Kmenta: visiting professor of economics, University of North Carolina, Chapel Hill, Jan.-Mar. 1973.

Charles B. Knapp, University of Wisconsin: assistant professor, department of economics, University of Texas at Austin.

Allen Kneese, Resources for the Future, Inc.: adjunct professor of economics, University of New Mexico, fall 1972.

John W. Knudsen: assistant professor of economics, University of Idaho, fall 1972.

Merphil S. Kondo: assistant professor of economics, University of Missouri-St. Louis, Aug. 30, 1972.

John E. Kwoka, Jr.: professor of economics, University of North Carolina, Chapel Hill, Aug. 1, 1972.

Sharon G. Levin: visiting instructor of economics, University of Missouri-St. Louis, Aug. 30, 1972.

Harold Gregg Lewis, University of Chicago: visiting professor, University of Georgia, fall 1972.

James J. McLain, University of Pittsburgh: assistant professor of economics and finance, Louisiana State University, Sept. 1972.

J. Ralph McLemore: instructor of economics, Missouri Southern State College, Aug. 23, 1971.

Lee R. McPheters, Virginia Polytechnic Institute and State University: assistant professor of economics, Florida Atlantic University, winter 1972.

Charles D. McQuillen: associate professor of business administration, Furman University.

Alex R. Maurizi: visiting associate professor of economics, University of North Carolina, Chapel Hill, 1972-73.

Wayne L. Miller: assistant professor, accounting and information systems, College of Business Administration, University of Nevada, Aug. 28, 1972.

Morris E. Morkre, California State College at Fullerton: senior lecturer in economics, University of Hong Kong, Sept. 1972.

Marnie W. Mueller: visiting assistant professor, department of economics, Wesleyan University, 1971-72.

R. Andrew Muller, University of Toronto: assistant professor of economics, McMaster University.

Martin Murphy: instructor, department of economics, Northeastern University, Sept. 1972.

Alfred L. Norman, University of Minnesota: assistant professor, department of economics, University of Texas at Austin.

Frank W. Oeschli: research demographer, acting assistant professor, Food Research Institute, Stanford University.

Claire L. Olsen, Northwestern University: assistant professor, department of economics, University of Miami, Sept. 1, 1972.

Margaret Oppenheimer: instructor, department of economics, Northeastern University, Sept. 1972.

Gerald T. O'Mara: assistant professor of economics, Northwestern University.

Barbara Pence: instructor, department of economics, College of Wooster, Sept. 1972.

Lydia M. Pitts, Columbia University: instructor, department of economics, City College of the City

University of New York, Sept. 1, 1972.

Alan I. Rapoport, University of Chicago: visiting lecturer, department of economics, University of Texas at Austin, 1972-73.

Howard M. Reed: lecturer in economics, College of Business Administration, University of Nevada, Aug. 28, 1972.

David Richardson: associate professor of economics, University of Kentucky, July 1, 1972.

Charles R. Roll, Harvard University: staff member, economics department, The Rand Corporation, Sept. 1972.

Don E. Roper, Federal Reserve Board: lecturer, department of economics, University of Texas at Austin, spring 1973.

Donald D. Rugg, University of California, Santa Barbara: staff member, management sciences department, The Rand Corporation, July 1972.

Donald F. Schaefer: post-doctoral fellow in economics, University of North Carolina, Chapel Hill, 1972-73.

Robert H. Schueler: lecturer, accounting and information systems, College of Business Administration, University of Nevada, Aug. 28, 1972.

Grant M. Scobie: instructor of economics, North Carolina State University, Aug. 1972.

Bethuel P. Setai, Essex County College: lecturer in economics, Merrill College, University of California, Santa Cruz, July 1, 1972-June 30, 1974.

Suresh P. Sethi, Stanford University: visiting assistant professor of management science, Rice University, July 1, 1972.

Richard E. Shaw, Northwestern University: assistant professor, department of economics, Wayne State University, Sept. 1972.

John J. Siegfried: assistant professor of economics, Vanderbilt University, Sept. 1, 1972.

Fred D. Sobering, North Dakota State University: extension professor, North Carolina State University, Sept. 1972.

Joseph J. Spengler: visiting professor of economics, University of North Carolina, Chapel Hill, 1972-73.

R. S. Stewart, Bluefield State College: associate professor, department of business and economics, Slippery Rock State College, Sept. 1972.

Wayne Thirsk, Prices and Incomes Commission, Canada: adjunct visiting assistant professor, department of economics and program of development studies, Rice University, July 1, 1972.

Daniel S. Tilley: research associate, department of economics, Iowa State University.

Terrie E. Troxel: lecturer, managerial sciences, College of Business Administration, University of Nevada, July 1, 1972.

E. Lane Vanderslice, Jr.: assistant professor of economics, Rutgers College, Feb. 1, 1972.

Claude M. Vaughan: visiting assistant professor of economics, University of Kentucky, July 1, 1972.

James F. Vetch: assistant professor of economics, University of Missouri-St. Louis, Aug. 30, 1972.

Larry E. Westphal: associate professor of economics, Northwestern University.

David Whipple, University of Kansas: assistant professor of economics and operations research, Naval

Postgraduate School, Sept. 1971.

Kenneth White, University of Wisconsin, Madison: assistant professor of economics, Rice University, July 1, 1972.

Ann D. Witte: visiting assistant professor of economics, University of North Carolina, Chapel Hill, 1972-73.

P. N. Worthington, Queen's College: associate professor, department of business and economics, Slippery Rock State College, Jan. 1972.

Neil R. Wright, Massachusetts Institute of Technology: assistant professor of economics, Rice University, July 1, 1972.

Uzi Yaari, Louisiana State University: assistant professor of economics, Rutgers College, 1972-73.

Julie H. Zalkind: visiting assistant professor of economics, University of North Carolina, Chapel Hill, 1972-73.

Gilroy J. Zuckerman: instructor of economics, North Carolina State University, Aug. 1972.

### *Leaves for Special Appointments*

Philip W. Bell, Merrill College, University of California, Santa Cruz: director, University of California Education Abroad Program, University of Nairobi, Kenya, 1972-74.

David W. Breneman, Amherst College: staff director, National Board on Graduate Education, National Research Council, June 1, 1972.

Kenneth T. Cann, Western Kentucky University: consultant, Universidad de Austral, Valdivia, Chile, July 1972.

Peter J. Cassimatis, Fairleigh Dickinson University: visiting research economist, Center of Planning and Economic Research, Athens, Greece, Sept. 1972-Sept. 1973.

George E. Delchanty, Northwestern University: chairman, department of economics, University of Nairobi, Kenya, Rockefeller Foundation, 1972-73.

Robert B. Desjardins, University of North Carolina, Chapel Hill: University of Utah MBA Program in Germany and England, Sept. 1, 1972-Aug. 31, 1973.

Richard B. DuBoff, Bryn Mawr College: visiting reader, Institute of Social Studies, The Hague, Netherlands, Sept. 1, 1972-July 30, 1973.

Robert E. Gallman, University of North Carolina, Chapel Hill: visiting faculty member, Nuffield College, Oxford, Sept. 1, 1972-Aug. 31, 1973.

George A. Hay, Yale University: special assistant to the Assistant Attorney General, Antitrust Division, U.S. Department of Justice, July 1972-July 1973.

William F. Hellmuth, McMaster University: International Monetary Fund Income Tax Mission to Kenya, Tanzania, and Uganda, spring 1972.

Ronald F. Hoffman, Social Security Administration: senior staff economist, Council of Economic Advisors, July 1972.

Ronald E. Kramer, Western Kentucky University: staff economist, U.S. Department of Commerce, Oct. 1971.

Benton F. Massell, Food Research Institute: program manager, Division of Social Systems and Human Re-

sources, National Science Foundation, 1972-73.

Leo V. Mayer, Iowa State University: senior staff economist, Council of Economic Advisors, Aug. 1, 1972-July 31, 1973.

Nicholas A. Michas, Loyola University: visiting professor, Athens Graduate School of Economics and Business Sciences; senior economist, Center of Planning and Economic Research, Athens, Greece, 1971-72.

Thomas I. Ribich, University of North Carolina, Chapel Hill: visiting faculty member, University of Wisconsin, Sept. 1, 1972-Aug. 31, 1973.

Charles E. Richter, University of North Carolina, Chapel Hill: Division of Urban and Regional Studies, Colombian National Department of Planning, Sept. 1, 1972-Aug. 31, 1973.

Jati K. Sengupta, Iowa State University: visiting professor, Indian Institute of Management, Calcutta, 1972-73.

Charles J. Stokes: supervisory economist, general accounting office, Sears Foundation, American Association of Collegiate Schools of Business Federal Faculty Fellowship, 1972-73.

James H. Street, Rutgers University: visiting scholar, Centre of Latin American Studies, Cambridge University, England, 1972-73.

Erik Thorbecke, Iowa State University: International Labor Office, World Employment Program, United Nations, 1972-73.

William P. Travis, University of California, San Diego: economic adviser to the Moroccan government, University of Michigan Center for Research in Economic Development, July 1, 1972.

Arthur B. Treadway, Northwestern University: Fulbright Fellow, La Universidad Autonoma de Madrid, Spain, 1972-73.

Robert C. Vogel, Southern Illinois University: senior staff economist, Council of Economic Advisors, July 1, 1972.

Samuel H. Williamson, University of Iowa: professor of economics, University of British Columbia, Sept. 1972.

### Resignations

Sydney N. Afriat, University of North Carolina Chapel Hill, Aug. 31, 1972.

Iftekhar Ahmed, Iowa State University: Bangladesh Planning Commission.

Larry D. Bedford, Iowa State University: Armour and Company, Nampa, Idaho.

Joseph R. Bisignano, Rutgers College: Federal Reserve Bank, San Francisco, June 1972.

H. James Boisseau, Jr., University of North Carolina Chapel Hill, Aug. 31, 1972.

Eugene A. Brady, Iowa State University: Indiana University.

Thomas K. Fitzgerald, Western Kentucky University, Aug. 1972.

George R. Iden, University of North Carolina Chapel Hill, Aug. 31, 1972.

Ferdinand K. Levy, Rice University: Georgia Institute of Technology, June 1, 1972.

William McFarland, University of North Carolina Chapel Hill, Aug. 31, 1972.

John M. Marshall, Vanderbilt University: University of California, Santa Barbara, Sept. 1, 1972.

Tridib K. Mukherjee, Appalachian State University: National Institute of Bank Management, Bombay, India.

James R. Prescott, Iowa State University: Temple University.

Larry E. Ruff, University of California, San Diego Environmental Protection Agency, Washington, D.C., June 30, 1972.

Raymond J. Struyk, Rice University: The Urban Institute, June 1, 1972.

Douglass W. Webbink, University of North Carolina Chapel Hill, Aug. 31, 1972.

Jerome D. Wiest, Rice University: University of Utah, July 1, 1972.

### Miscellaneous

Ronald A. Krieger, economics editor, *Business Week*.

### NOTE TO DEPARTMENTAL SECRETARIES AND EXECUTIVE OFFICERS

When sending information to the *Review* for inclusion in the Notes Section, please use the following style:

A. Please use the following categories:

- |   |   |
|---|---|
| 1—Deaths  | 6—New Appointments                                  |
| 2—Retirements                                   | 7—Leaves for Special Appointments (NOT Sabbaticals) |
| 3—Foreign Scholars (visiting the USA or Canada) | 8—Resignations                                      |
| 4—Promotions                                    | 9—Miscellaneous                                     |
| 5—Administrative Appointments                   |   |

B. Please give the name of the individual (SMITH, John W.) his present place of employment or enrollment: his new title (if any), his next place of employment (if known or if changed), and the date at which the change will occur.

C. Type each item on a separate 3x5 card, and please do not send public relations releases.

D. The closing dates for each issue are as follows: *March*, November 1; *June*, February 1; *September*, May 1; *December*, August 1.

This announcement supersedes and replaces a letter which was sent annually from the managing editor's office. All items and information should be sent to the Assistant Editor, American Economic Review, Box Q, Brown University, Providence, Rhode Island 02912.

## PROPOSED WASHINGTON NEWSLETTER

The Committee on Policy Communications of the American Economic Association has been asked by the Executive Committee to explore the extent to which Association members might be interested in a proposed new "Washington Newsletter." The newsletter would be designed primarily to provide "early warnings" about on-going or planned research projects, upcoming Congressional hearings and other Washington activities of interest to economists; it would not take policy positions. It has not yet been determined under whose auspices the newsletter might be published. A prototype of such a newsletter is presented below.

The Committee would be most interested in hearing your reaction to the prototype. In reading it over, the following points should be kept in mind:

1. The prototype was prepared in February 1972 and has not been updated.
2. The prototype is mainly designed to indicate the type of material to be included in a newsletter; if the decision were made to go ahead with publication of such a letter, further efforts would be made to improve its style and readability.
3. The prototype does not give comprehensive coverage of Washington developments, just as a single issue of the proposed biweekly newsletter would not. The aim in each issue would be to report on the most important developments and to cover a limited number of other areas in a more systematic fashion. The latter activity should insure that twenty-six consecutive issues of the newsletter would provide relatively full information about Washington activities of interest to economists that are not adequately covered by other sources.

Specifically, the committee would like your reactions to the following questions:

1. What is an example of a good report and why?
2. What is an example of a poor report and why?
3. Are there other topical areas that should be covered?
4. Should some types of information be given greater coverage; for example, listings of upcoming publications and information about employment opportunities?
5. Would you be willing to pay an annual subscription price of \$10 for such a biweekly newsletter? If it proved to be necessary, would you be willing to pay \$15?

Reactions should be sent to Rendigs Fels, Secretary-Treasurer, AEA, 1313 21st Avenue, South, Nashville, Tennessee 37212.

## PROTOTYPE OF PROPOSED WASHINGTON NEWSLETTER

Vol. 0, Number 0

March 1972

### JEC Schedules March Hearings on Value-Added Tax; Ways and Means Committee May Deal with Tax Reform

Hearings on the value-added tax will be held by the Joint Economic Committee, Mar. 21-24. Among those to appear before the full committee are Richard Musgrave, Harvard University; Sheldon Cohen, former commissioner of Internal Revenue; Henry Aaron, Brookings Institution; Robert Lampman, University of Wisconsin, and Lester Thurow, Massachusetts Institute of Technology. No witnesses representing the Administration have been scheduled. Hearings outside Washington are also anticipated by the committee staff.

At the request of President Nixon, the Advisory Commission on Intergovernmental Relations will analyze the impact of the value-added tax on state/local property taxes. The study, directed by ACIR's John Shannon, is not expected to result in a formal legislative proposal by the Administration during 1972. The Treasury Department's Office of Tax Analysis will also be involved in the White House-initiated examination of the tax.

*Related tax developments:* Widespread discussion of the value-added tax as well as the continuing budget deficit has given support to moves for broad tax reform, notably in the federal income tax, and for state tax effort provisions in several major revenue-sharing proposals. Despite pressures from the Democratic Study Group in the House, it is uncertain whether Rep. Wilbur Mills (D-Ark.) will schedule Ways and Means Committee hearings on tax reform. Although skeptical that significant legislation could pass this year, Mr. Mills may permit tax reform to be a part of hearings on the Administration's request for an increase in the debt ceiling to \$480 billion.

Should Ways and Means take up tax legislation, two principal measures to be considered would be those sponsored by Reps. James Corman (D-Calif.) and Henry Reuss (D-Wis.). The Corman bill, drafted by the staff of the Joint Committee on Internal Revenue Taxation, is an extensive listing of tax provisions prepared chiefly for discussion purposes. The Reuss bill represents the consensus of the Democratic Study Group.

Revenue gains envisioned by the Corman bill were estimated by the Joint Committee staff as \$11 billion in calendar year 1973 and \$19 billion when fully effective in 1980. These figures do not reflect the impact of the repeal of the investment credit and of the new depreciation range. Reuss estimated that his bill's provisions would result in new revenue gains of \$7.5 billion.

In the Senate, a bill sponsored by Sen. Gaylord Nelson (D-Wis.) will be the basis of discussion on the issue, although no hearings are presently scheduled in the Senate Finance Committee. The measure also draws

on the work of the staff of the Joint Committee on Internal Revenue Taxation, although it contains additional input from the Reuss bill.

### JEC Papers Examine Inflation, Subsidies, and Welfare Programs

The Joint Economic Committee has undertaken major research efforts on welfare and federal subsidy programs. Hearings on these and other issues will be held during the spring. (See article on Congress.)

Congress in 1971 appropriated \$500,000 at the request of Rep. Martha Griffiths (D-Mich.), chairman of the Subcommittee on Fiscal Policy, for a comprehensive two-year study of the nation's system of welfare-related programs. Total federal, state, local, and private welfare costs now total over \$100 billion.

The welfare study will consider: 1) who and what kinds of people are now covered; 2) how welfare recipients differ from noneligibles; 3) why welfare rolls are expanding so rapidly; 4) how programs integrated as to benefit levels, administration, and work levels; 5) the relationship between welfare, "employability," and the structure of the labor market; 6) how much a family would have to earn after taxes to equal the package of goods and services provided by welfare; and 7) what the effects of various welfare programs are on family structure, family responsibilities, and the need for continuing on welfare? Interim reports will soon be available. The overall study is due by June 30, 1973, and is directed by Ms. Alair A. Townsend under the supervision of James W. Knowles, director of research for the JEC. Inquiries can be made to the study office in Room 1537, Longworth Building, Washington, D.C. 20515; telephone (202) 225-3565.

As a follow-up to the January hearings on subsidies by Senator William Proxmire's (D-Wis.) Subcommittee on Priorities and Economy in Government, approximately 40 study papers by academic and other experts will be published by the JEC. These papers will cover general subsidy issues and 11 specific areas: agriculture, food, education, manpower, international, housing, natural resources, transportation, tax subsidies, commercial television broadcasting and stockpiling of strategic materials. An initial study, "The Economics of the Federal Subsidy Program," prepared by Jerry J. Jasinowski of the JEC staff and Dr. Carl S. Shoup was published Jan. 11, 1972 and is available from the JEC.

Two specialists' papers were also commissioned by JEC: Otto Eckstein, "The Inflation Process in the United States"; and Lester Thurow, "The American Distribution of Income: A Structural Problem." Eckstein's paper is available, and Thurow's will be published soon.

### Study on Motor Carriers Ready for ICC by June

The Regular Common Carrier Conference of the American Trucking Association expects to submit in June a major study to the Interstate Commerce Commission. The year-long study, commissioned for \$210,000, is designed to produce data on "the sum of money needed

(by regular carriers) to provide financial stability, to attract debt and equity capital . . ."

The study, directed by consultant Irwin H. Silberman, formerly professor of banking and economics at New York University, was begun in 1971 as an industry response to ICC's mandate to assure financial stability of the carriers. The study's intent is to develop methods of risk measurement applicable to the industry for the use of the ICC and company management.

Upon its submission to ICC, the study will become part of rule-making proceedings, particularly MC 82 now pending, which require carriers to introduce evidence to support requests for rate increases. MC 82 is expected to be a precedent-setting decision. Initiated by ICC, it is an examination of the basic methods by which ICC reviews rate increases.

The RCCC, sponsor of the Silberman study, is developing a capability to serve as a clearinghouse for data of interest to transport economists. Several summaries of other studies are available from the organization, located at 1616 P Street, N.W. Washington, D.C. 20009.

#### **Water Bill Allows NRC up to \$15 million for Impact Studies**

The House Public Works Committee is currently making final revisions in HR 11896, "The Water Pollution Control Act Amendments of 1971." Passage of a major bill is expected before Congress adjourns. It is likely that the final version will authorize \$15 million for comprehensive studies sponsored by the National Academies of Engineering and Sciences into "all aspects of the total social and economic effects of achieving or not achieving" the national effluent objectives for 1981. The studies, to be coordinated by the National Research Council, would be due—along with recommendations—two years after the enactment of the bill.

Several other provisions of the House bill would involve major studies of economic factors:

1. *Thermal discharges.* The Environmental Protection Agency would be required to conduct comprehensive studies of costs, benefits, impact and technology involved in thermal discharges. This report would be due in one year.
2. *Foreign nations.* The Commerce Department would study the effect that the costs of undertaking abatement control programs would have on the competitive position of U.S. manufacturers as compared with foreign manufacturers not required to comply with as stringent standards. The study would include alternative means of equalizing competitive advantages. The bill authorizes \$1 million for the study. An initial report would be due in six months.
3. *National Policy.* The President would be required to investigate and report within two years on all national policies and goals involving environmental quality. The evaluation, for which \$5 million would be authorized, would take into account the nation's available resources.

In addition to these specific directives, the bill's thrust toward stringent standards and stiff compliance would require extensive consideration by companies of

the economic and social costs of designing their new and existing facilities, since the bill contemplates relief from standards where costs outweigh benefits.

In summary, the 1972 water pollution bill will confront the economic impact of stringent pollution controls by authorizing \$15 million in studies to be contracted by the National Research Council.

#### **New-Left Think Tank Focuses on Neighborhoods**

The Institute for Neighborhood Studies has been recently organized as a nonprofit research, teaching, and action foundation in Washington, D.C. Founded by Milton Kotler, a resident fellow of the Institute for Policy Studies, the Institute for Neighborhood Studies as a "think tank" for the New Left focuses on a radical approach to neighborhood organization and neighborhood control.

Two initial research interests of the INS center on the urban public finances of Washington, D.C. John Delaplaine heads one project measuring the net flow of D.C. public sector spending to suburbanites. Another project has partly served as the intellectual infrastructure for a Neighborhood Corporation funding bill introduced by Sen. Mark Hatfield (R-Ore.). Preliminary results of this project indicate that the "visible benefits" of government in a low income area of the central city (Shaw-Cordozo) actually are less than the total tax payout from the area, thus providing an economic rationale for greater neighborhood self-government and self-finance. Along with Kotler, other key fellows of INS include Gerson Green, former research director for Demonstration Projects of the Office of Economic Opportunity, and Frank Smith. The INS is located at 1520 New Hampshire Avenue, N.W., Washington, D.C. 20036; telephone (202) 234-9382.

*Related development:* A separate research effort on the net flow of public sector benefits to suburbanites is underway at the Urban Institute. For further information call Claudia Scott (202) 223-1950.

#### **Productivity in Government is Topic of Studies by Urban Institute, Inter-Agency Task Force**

Two separate studies dealing with the issue of productivity in government—federal, state, and local—are under way in Washington. The National Commission on Productivity is supporting investigations by the Urban Institute, while the Office of Management and Budget is coordinating an inter-agency task force which is considering federal government productivity indices.

*Urban Institute:* Under a \$75,000 contract with the Productivity Commission, the Institute expects to submit a report by June 1 which will provide two types of information:

1. The feasibility of developing improved measurements of productivity in local government for solid waste collection and police protection. These functions were chosen as representative of major cost items in local budgets. The study seeks to answer what data can be used both intra-jurisdiction and inter-city.

2. The benefits and transferability of previous innovations in solid waste collection and law enforcement. With the cooperation of the International City Management Association, the Institute identified a number of possible examples of increased productivity in these fields. They will be systematically evaluated by the Institute.

The Productivity Commission's preliminary work program for 1972-73, issued Feb. 18, indicates that support for this Urban Institute research program will be continued in the future. The study is directed by Harry Hatry, formerly with the Office of Management and Budget.

*Task Force:* The OMB, Civil Service Commission, and General Accounting Office, with some staff support by the Bureau of Labor Statistics, are participating in studies to test the feasibility of developing productivity indices covering about 50 percent of the federal civilian work force. Completion of the initial studies is expected by June 30.

The project has four points of emphasis:

1. Development of productivity indices for broad functional responsibilities (e.g., administrative services, service to the public) rather than agency-by-agency or program-related.
2. Uses of measurement data by managers. This element of the study includes identification and analysis of incentives and disincentives for managers to seek greater productivity.
3. Uses of cost data by managers. The absence of cost data in the federal government is being weighed along with possible reasons for the failure to use such data where available. The study will attempt to show how unit cost information and technical methodology can improve justifications for capital investment proposals.
4. Development of effectiveness measurements in service establishments. The study seeks to move beyond customary criteria of efficiency to weigh impacts and quality of services.

After approval by the member agencies' principals, recommendations of the study may be made public. It is expected that additional projects will be conducted in fiscal 1973.

*Related development:* In a more operational sense, the productivity of the federal government is also the province of the interagency Regulations and Purchasing Review Board. Its Feb. 1 progress report documents studies by the board and notes its recommendations for executive and legislative action. Copies may be obtained from Dr. Jack W. Carlson, executive director, Regulations and Purchasing Review Board, Room 241, Old Executive Office Building, Washington, D.C. 20503.

#### **Task Force to Submit Report on Unemployment Statistics to Nixon**

The Administration's Task Force on Employment and Unemployment, chaired by Treasury Under Secretary

Charles E. Walker, is expected to present its recommendations to President Nixon in the near future. The interagency study group, which was appointed Jan. 20, has considered the following matters:

1. The soundness of present statistical methods for compiling unemployment data and the consistency of the resulting figures.
2. The age, sex, educational, and geographic characteristics of the unemployed and how they differ from earlier periods.
3. How long it takes a discharged worker to find a new job.
4. Whether unemployment compensation, welfare payments and minimum wage legislation are disincentives to increasing employment.

Treasury Secretary John B. Connally also asked members to recommend "changes in policy or additional policy" that are needed to deal with the current situation.

The view that there has been a structural change in the work force, expressed in congressional testimony by the Council of Economic Advisers and other administration officials, could lead the task force to recommend a redefinition of the full employment level. The target for maximum unemployment has been about 4 percent since 1962.

Members of the task force are: Mr. Walker; Edgar R. Fiedler, assistant Treasury secretary for economic policy; James T. Lynn, Commerce Department under secretary; Ezra Solomon, member, Council of Economic Advisers; Michael H. Moskow, Labor Department assistant secretary; Laurence E. Lynn, HEW Department assistant secretary; Carl W. Clewlow, Defense Department deputy assistant secretary; and Don A. Paarlberg, director of agricultural economics, Agriculture Department.

#### **New Tax Reform Group Receives Nader Funds**

Tax Reform Research Group was established Jan. 1, 1972, funded by Ralph Nader's nonprofit Public Citizens, Inc. Headed by Harvard Law School graduate, Thomas H. Stanton, the Research Group intends to build a team of lawyers and economists to research and act in the tax policy arena.

In 1971, some members of this group challenged the Treasury's Asset Depreciation Range (ADR) regulations by both law suits and public debate. Now the group is beginning analysis of the tax policy process in order to recommend and work for tax reforms. Initial interest is in the federal income tax, state and local taxation, especially the property tax, and in developing citizen awareness of the Administration's proposed value-added tax as a property tax partial substitute.

The Group is searching for young economists interested in public finance and tax reform. Those interested may contact Thomas H. Stanton, Tax Reform Research Group, Suite 426, 733 15th Street, N.W., Washington, D.C. 20005; telephone (202) 783-6840.

### Concentrated Industry Studies Move in FTC

Extensive investigations begun in 1968 by staff of the Federal Trade Commission in "concentrated industries" are nearing completion for submission to the members of FTC. Four industries—electrical equipment, energy, automobiles, and ethical drugs—are the subjects of the studies which are being coordinated by Dr. Michael Mann of FTC. The fiscal 1972 budget for the studies is approximately \$275,000, divided fairly evenly among the four.

Furthest advanced are studies on (1) whether there is a single fuel market and (2) whether total fuel sources have been curtailed as a result of oil company acquisitions of other sources. No target date has been set for the ethical drug study.

### Productivity Commission Lists Agenda for 1973; To Sponsor Research in Construction, Health

The first annual report of the National Productivity Commission was issued Mar. 10 by its outgoing chairman George P. Shultz, director of the Office of Management Budget. New chairman of the commission is Secretary of Commerce Peter Peterson. Of particular interest are the commission's future plans, since the Administration is requesting \$5 million to support the group's objectives as established by the Economic Stabilization Act of 1971.

Under this legislative mandate, the commission will extend its activities in several directions. It plans to expand its substantive program of policy research and development. Most of this basic work will be done by the commission staff. They will enlist the aid of outside consultants, provide background papers, arrange for seminars and conferences, and draft recommendations for the commission.

Special attention will be given to work on the opportunities and obstacles to improving productivity in important industries which have been lagging. In this effort, the commission will strive to be responsive to the needs and suggestions of the agencies involved in the Economic Stabilization Program. In this context, the studies of government productivity will be continued. Also, work will be undertaken on several other key industries, including construction and health, where costs have been rising sharply.

Another broad area of examination by the commission will involve factors that affect productivity generally. Alternative approaches to pollution control will be analyzed in terms of both relative efficiency in achieving goals and implications for productivity growth. Studies will be made of the influence on productivity of worker attitudes and motivation, and the possibilities of improvement through group incentive plans, job redesign and related techniques. The contribution of R & D programs to productivity enhancement will also be covered. These areas will be explored with the assistance of leading experts in government, universities, and private industry who are already giving attention to these problems. To give attention to productivity improvement on a regional and local basis, the regional representative of the Secretary of Labor will be given responsibility for

development of a coordinated federal effort to promote greater awareness of productivity at the local level and in specific industries at the regional level. These representatives will receive technical assistance from the commission's National Office and will draw on the work of the commission to promote the objectives of the 1971 Act. With the Regional Councils as their base, the regional representatives will be able to utilize the resources of a wide range of domestic departments and federal agencies who have close working relationships with state and local governments and broad contact with all sectors of the public.

The commission will also enlarge its information program to improve public understanding of the issues and its program. Workers, managers, and consumers will be given more information about the meaning and importance of productivity and its relationship to jobs, income and living standards. The commission's studies, reports and recommendations will be disseminated widely. National, regional, and local conferences, meetings, workshops and seminars with a wide range of participants will be held. Special pamphlets, speeches, and other educational materials will continue to develop recommendations for private and public policy to encourage productivity improvement in the decade of the 1970's.

### Criteria to Evaluate Public Works Projects Inspire Controversy

The Water Resources Council has issued proposed standards for planning water and land resources, which set new criteria for analyzing the merits of future dams, barge canals, and other federal water development projects. Noting that past decisions have been based primarily on monetary considerations, the Council said its new principles and standards for water and land resource planning adopt a multiobjective approach, giving full consideration to national economic development, environmental quality and regional development.

Scheduled for implementation in late spring 1972, the standards are the result of a two-year study and review of decision-making practices undertaken by a special Council Task Force. There will be a regionalized approach to implementation of land and water resource programs and the new standards will require development of alternative plans for each project.

Conservationists claim that, while the proposed standards set a discount rate (7 percent for the next five years rather than the present 5 percent range) which provides a better measure of the economic cost of these projects, and include certain environmental considerations, they also appear to give public works agencies like the Corps of Engineers and the Bureau of Reclamation excessive latitude in making cost-benefit evaluations. Sens. Jennings Randolph (D-W. Va.) and Henry M. Jackson (D-Wash.) have introduced legislation (S 2612) that would adopt the new "multi-objective" approach but cut the discount rate to the 3 percent range.

No further research will be supported by the Water Resources Council on the proposed standards which were published in the *Federal Register*, Dec. 21, 1971. Public hearings will be held during March. If one wishes

to submit comments, they should be sent by Mar. 31 to W. Don Maughan, director, Water Resources Council, 2120 L Street, N.W., Washington, D. C. 20037; telephone (202) 254-6303.

### Materials Policy Group Seeks Academics' View

The National Commission on Materials Policy, authorized by the National Materials Policy Act of 1970, must report to the President and Congress by June 30, 1973. Although its fiscal 1972 budget is \$500,000, the commission has been hampered by slowness in hiring staff and setting its priorities. Most of its research will be conducted by outside firms and consultants.

Staff of the commission indicate a pressing need for several broad-gauged economists who can help set the research agenda. The staff is also highly interested in views of the academic community. Anyone interested in providing input should contact Dr. James Boyd, executive director, or Ralph E. Burr, deputy executive director, at the commission, 1016 16th Street, N.W., Washington, D. C. 20036; telephone (202) 382-4735.

### NBS Assesses Economics of Factory-Built Housing

Under an interagency agreement with the Department of Housing and Urban Development, the National Bureau of Standards is conducting an economic analysis of factory-built housing systems as part of "Operation Breakthrough," HUD's effort to stimulate factory-built housing. The estimated \$500,000 study seeks to identify the scale of production at which sufficient economies can be achieved to reduce housing costs.

In order to supplement its staff for the study, NBS is seeking a senior economist with building construction background. Initial contact should be made with Dr. Howard Morgan, National Bureau of Standards, Washington, D.C. 20234; telephone (301) 921-3135.

### FCC to Pursue Extended Study of AT&T Rate Request

House Appropriations Subcommittee hearings Mar. 14 will reveal the plans FCC has to pursue its once-dropped long-term investigation of AT&T's controversial request for long-distance rate increases. Officials of the Common Carrier Bureau, which would conduct the study for the commission, have estimated that two dozen additional professional staff members will be needed for a "reasonably creditable job." The subcommittee will consider bills sponsored by several congressmen which would assure funding of staff increases for this purpose.

Remaining questions to be answered by FCC to determine whether AT&T's rate requests would be lawful are:

1. What are the interstate revenue requirements to keep the company in operation, and on what basis are these requirements to be determined?
2. How much operating revenue can be expected to accrue from the new rates?
3. Whether the differing rates proposed by AT&T for the various types of long-distance calls are justified in terms of cost to the company, demand by customers, and other factors.
4. Whether the proposed rates, in light of these determinations are just and reasonable as required by law.
5. Whether the FCC should prescribe just and reasonable, or maximum and/or minimum charges, for long-distance calls, and, if so, what those charges should be.

Bernard Strassburg, chief of the commission's Common Carrier Bureau, has said that it could be two years before these issues are actually before the FCC for a final decision.

### Congressional Hearings Schedule

*Joint Economic Committee* (full): Mar. 21-24, value-added tax; Apr. 19-22, productivity; Apr. 24-26, wage-price controls.

*Subcommittee on Fiscal Policy*: End of March, welfare reform; subsequent field hearings include New York, (Apr. 11-13), Detroit (May 3-5) and Atlanta (tentatively June 6-8). *Subcommittee on Priorities and Economy in Government*: occasional hearings on federal subsidies throughout 1972.

### Senate

*Judiciary Subcommittee on Antitrust and Monopoly*: Through April 15, housing loans to minorities, sales of revenue bonds by commercial banks, controls and competition under Phase 2, selling below cost. After Apr. 15: hospital costs, corporate farming, wage-price controls, energy issues, gasoline marketing, and possibly no-fault insurance.

*Interior and Insular Affairs*: Through Apr. 13: timber management, Department of Interior energy study, outer continental shelf policy.

*Commerce Subcommittee on Environment*: Through March: solid waste management and packaging. *Subcommittee on Surface Transportation*: continuing hearings on deregulation of common carriers.

*Foreign Relations*: Apr. 17-20, military aid.

*Ways and Means*: Through end of March: markup of revenue sharing bill; April, possible markup of health insurance, possible hearings on taxes.

*Judiciary Subcommittee #5 (antitrust)*: Not scheduled but anticipated hearings on corporate farming, corporate control act.

*Banking and Currency*: Mar. 28: value of gold and the dollar.

*Merchant Marine and Fisheries*: Late March-April: changes in the Environmental Policy Act. Hearings anticipated on trans-Alaska pipeline after Department of Interior releases its environmental impact statement.

## SIXTY-NINTH LIST OF DOCTORAL DISSERTATIONS IN POLITICAL ECONOMY IN AMERICAN UNIVERSITIES AND COLLEGES

The present list specifies doctoral degrees conferred during the academic year terminating June 1972. Abstracts of many of the dissertations are supplied.

### General Economics; including Economic Theory, History of Thought, Method- ology, Economic History, and Economic Systems

MANUEL R. AGOSIN, Ph.D. Columbia 1972. Incentive systems for socialist firms.

This study analyzes several aspects of managerial incentive systems in socialist economies. It shows the specific characteristics of socialist markets and draws the implications of the various incentive systems for the possibilities of utilizing a decentralized model of socialism. A hypothesis regarding the effects of different incentive systems on economic growth is put forth, and evidence in its favor is obtained from the experience of Yugoslav industry during the 1954-64 period.

KARL ASMUS, Ph.D. Michigan State 1972. The choice of an optimal consumer planning horizon.

RICHARD N. BEAN, Ph.D. Washington 1971. The British trans-Atlantic slave trade: 1650-1775.

BARRY P. BERLIN, Ph.D. Duke 1972. An economic analysis of foundation grants.

ERNST R. BERNDT, Ph.D. Wisconsin (Madison) 1972. The economic theory of separability, substitution, and aggregation with an application to U.S. manufacturing: 1929-68.

In the context of production theory, it is shown that functional separability, consistent aggregation, the path independence of Divisia indexes, and certain equality restrictions on the Allen partial elasticities of substitution are equivalent restrictions on any strictly quasi-concave homothetic production function. I then perform two empirical studies which employ the translog production function and data from U.S. manufacturing; I test for the separability and aggregation conditions by imposing parameter restrictions on the translog function.

GEORGE C. BITROS, Ph.D. New York 1972. Replacement of the durable inputs of production: A theoretical and empirical investigation.

Contemporary studies of the demand for fixed (capital) and quasi-fixed (labor) inputs sidestep the problem of replacement by assuming that depreciation is a rigid engineering parameter. This dissertation uses the neoclassical theory of the firm to develop a model within which replacement is subject to entrepreneurial choice. Depreciation is due to utilization and embodied

technological change, both of which are treated as decision variables. Finally, the capital and labor replacement equations are employed to explain a set of data obtained from a large telephone utility.

DANIEL R. BLAKE, Ph.D. Oregon 1971. Separable utility functions with an application to consumption-work choices.

JEAN-MARIE BLIN, Ph.D. Purdue 1972. Social decision process.

This study starts from the general concept of an individual preference pattern and its various mathematical representative structures to attack the aggregation problem. Some randomized strategy properties of majority voting are first studied in the context of constitutional choice. The concept of a collective decision rule as a pattern classifier is then used to characterize PO states in a public goods economy. The mathematical foundations of aggregation theory are presented and used to derive various aggregation algorithms possessing some desirable properties (for example, transitivity, completeness, etc.).

ALAN S. BLINDER, Ph.D. Massachusetts Institute of Technology. Towards an economic theory of income distribution.

VOLKER H. BOEHM, Ph.D. California (Berkeley) 1972. Stable firm structures and the core of an economy with production.

JOHN BROOME, Ph.D. Massachusetts Institute of Technology. Economics with indivisible commodities.

BARBARA A. BURSTEIN, Ph.D. Wayne State 1971. A two-sector growth model: Stability criteria.

DONALD E. CAMPBELL, Ph.D. Princeton 1972. Voting and social choice.

A collective choice rule yielding transitive social orderings is shown to satisfy seven criteria which imply Arrow's five conditions *in practice*. A social choice algorithm, constructed from simple majority rule, selects an alternative which is socially best according to any collective choice rule satisfying the seven conditions. The algorithm and some progressive taxation criteria are used to determine the ideal distribution of wealth which assigns an equal share of national wealth to each individual.

RICHARD J. CEBULA, Ph.D. Georgia State 1971. The commodity trap.

This study provides a theoretical framework which may help to explain why monetary policy, even though it affects the rate of interest, may be unable to influence aggregate spending, the aggregate level of output, or the aggregate level of employment in an economy. The conceptual framework developed in this dissertation is referred to as the commodity trap.

DAVID S. C. CHU, Ph.D. Yale 1972. The Great Depression and industrialization in Latin America: Response to relative price incentive in Argentina and Colombia, 1920-45.

JOHN Y. COFFMAN, Ph.D. Purdue 1972. Some non-convex quadratic optimization techniques and their application to selected economic problems.

The dissertation presents two algorithms for the minimization of a nonconvex quadratic objective function subject to linear constraints. It then briefly discusses the application of nonconvex quadratic programming to two problems in economics: the two-person, nonzero-sum game, and the transportation problem with non-linear, concave cost functions.

PETER B. DIXON, Ph.D. Harvard 1972. A theory of joint maximization.

Equilibrium solutions in general equilibrium models may be found by maximizing the value of a "suitably" weighted sum of individual utilities subject to the combined production possibilities of all individuals. This idea is exploited in two directions. First, there is a discussion of algorithms for finding the suitable weights and hence the general equilibrium solution. Second, the joint utility function is used in an analysis of the problem of aggregation across households in demand theory.

ALAN I. DUCHAN, Ph.D. Wisconsin (Madison) 1972. Decision making in a macro-economic context.

DENNIS F. ELLIS, Ph.D. Wayne State 1972. The optimal allocation of risk and the demand for securities in an exchange economy.

PHILIP H. EMPEY, Ph.D. Purdue 1972. A stochastic revealed preference approach to the empirical investigation of the axiom of consumer consistency.

The micro-economic assertion that changes in one's purchasing pattern are based upon changes in either income or market prices is investigated by applying the Strong Axiom of Revealed Preference directly to a set of weekly expenditure data. The postulated stochastic choice model assumes that the observed purchases randomly deviate from the optimum preference bundles, and the nature of these observed deviations is examined. It is concluded that little inconsistency is actually revealed by the consumer.

ALLAN M. FELDMAN, Ph.D. Johns Hopkins 1972. Non-recontracting, recontracting, and equitable trading processes.

This dissertation is about the stability of barter trading processes in a pure exchange economy. First, it is shown that under certain conditions a sequence of bilateral trades leads to a pairwise optimum, and, under additional conditions, a pairwise optimum is efficient. Then stability theorems are established for random Paretian reallocation processes. Finally, a simple proof for Edgeworthian recontracting stability is established, and the analysis is extended to treat notions of equity, or fairness.

FRANCIS P. FERGUSON, Ph.D. Wisconsin (Madison) 1971. Analysis, interpretation, and institutional change: Marx and the institutionalists, Commons and Veblen.

WILLIAM D. FINKLE, Ph.D. Massachusetts Institute of Technology. Duality in economic models.

COLIN A. GANNON, Ph.D. Pennsylvania 1970. Contributions to the economic theory of spatial competition.

Locational equilibrium configurations for a class of spatial duopoly industry structures, consisting of a static, nonstochastic, bounded, one-dimensional spatial market, are identified under a rich variety of weak assumptions regarding conjectural interdependence, degree of product differentiation, pricing policies, and objectives of the firms. The tendency of firms to concentrate at the center of the market is found to be remarkably pervasive and not simply an aberration induced by perfectly inelastic product demand as suggested by the models of Hotelling and Smithies.

CHARLES C. GILLETTE, Ph.D. Oklahoma State 1972. The political economy of John Ruskin.

This study is an examination of Ruskin's criticism of classical economics, his attempt to reconstruct political economy, his proposals for economic reform, and his influence. Although it does not appear that any significant economists except Hobson were influenced by Ruskin, it is probable that he was influential in preparing the way for socialism and the welfare state in Great Britain.

JAMES R. GOLDEN, Ph.D. Harvard 1972. Investment behavior by United States railroads: 1870-1914.

The dissertation explores variations in investment behavior for different capital components, regions, and time periods by comparing estimated parameters of models drawn from current investment theory with qualitative evidence. Regression results of revised aggregate sectoral data were compared with corresponding findings for cross-section and time-series analysis of a nineteen company sample. In each case the aggregate data produced misleading results which diverged from the qualitative evidence and company conclusions because of significant regional behavioral disparities.

GEORGE W. GRANTHAM, Ph.D. Yale 1972. Technical and organizational change in French agriculture be-

tween 1840 and 1880: An economic interpretation.

JERRY W. GUSTAFSON, Ph.D. Johns Hopkins 1972. An essay on the concept of collective rationality.

RICHARD C. HARTMAN, Ph.D. California (Berkeley) 1971. A firm's investment decision with uncertain future prices.

LARENCE A. HIRSCHHORN, Ph.D. Massachusetts Institute of Technology. Toward a political economy of information-capital.

GLENN R. HUECKEL, Ph.D. Wisconsin (Madison) 1972. The Napoleonic Wars and their impact on factor returns and output growth in England, 1793-1815.

KATSUHIITO IWAI, Ph.D. Massachusetts Institute of Technology. Essays on dynamic economic theory of optimal capital accumulation and Keynesian short-run disequilibrium dynamics.

KUNIO KAWAMATA, Ph.D. Minnesota 1972. Price distortion and potential welfare.

A multi-commodity model is studied in which one group of agents is guided by different prices from those of the other group. It is established that a decrease in the specified divergence between the equilibrium price vectors implies an increase in potential welfare when the change in distortion is proportional. This broadens a conclusion of the classical theorem of welfare economics and answers a question of Foster and Sonnenschein in *Econometrica*, Vol. 38.

RICHARD H. KEEHN, Ph.D. Wisconsin (Madison) 1972. Market structure and bank performance: Wisconsin, 1870-1900.

JOHN H. KEITH, Ph.D. California (Berkeley) 1972. Inventory investment, the generalized flexible accelerator, and interrelated factor demand.

ELMO A. KELLER, JR., Ph.D. Iowa State 1972. The computation of optimal growth in economic models. Control theory algorithms, conjugate gradient and Davison methods are applied to growth models, using penalty functions for handling terminal state constraints. It is shown how the optimal time paths change with respect to changes in certain structural features of the economic growth model. The analysis is in the form of numerical experimentation with non-linear models under various economic hypotheses. There is also an examination of simulated optimization using feedback relationships between the state and control variables.

EDITH B. LANG, Ph.D. Rochester 1972. The effects of net interregional migration on agricultural income growth: The United States, 1850-60.

HARVEY E. LAPEN, Ph.D. Massachusetts Institute of Technology. Models of non-steady-state economic growth and a dynamic model of the firm.

THOMAS M. LENARD, Ph.D. Brown 1972. Aggregate policy controls and optimal growth: Theory and applications to the U.S. economy.

The dissertation explores the theoretical and empirical properties of the single sector neoclassical optimal growth model with labor augmenting technical change and a variation of that model in which saving is a function of employment.

EDWARD P. LEVEEN, Ph.D. Chicago 1971. British slave trade suppression policies, 1821-65: Impact and implications.

Two impacts of suppression policies are examined. First, it is argued that the navy's impact on the costs of slaving (considering also demand and supply elasticities) significantly reduced the volume of the illegal trade. Second, the impact of suppression upon various British interests is analyzed and it is determined that the naive model of economic self-interest advanced by Eric Williams cannot explain British determination to bring about an end to slave trading.

DILIP B. MADAN, Ph.D. Maryland 1972. Competitive equilibrium, survival, and social choice.

JACOB METZER, Ph.D. Chicago 1972. Some economic aspects of railroad development in Tsarist Russia.

The direct contribution of the railroads to economic growth in Tsarist Russia is assessed by estimating the saving of resources from their freight and passenger transportation in 1907. The share of the estimated saving in Russian GNP was 5.5 percent which indicates that the direct effect of the railroads on Russian economic progress was small. In addition, the evolution of a national grain market is quantitatively demonstrated and its link with the railroads expansion empirically established.

STEPHEN P. MEZGER, Ph.D. Rice 1972. An integrated decision model of the firm.

The purpose of this thesis is to derive decision rules for the firm, integrating variables of production and finance. A static model is developed to obtain rules for production and financial mix. Adaptations are made to managerial objectives. Dividend policy as a substantive decision must consider the possibility of future period investment opportunities. The model becomes dynamic and the tools provided by optimal control theory are used to derive the appropriate decision rules.

ROBERT J. MICHAELS, Ph.D. California (Los Angeles). Explorations in a two-sector vintage model of economic growth.

In this dissertation, a two-sector vintage growth model with differential possibilities for input substitution *ex ante* and *ex post* is constructed, analyzed, and

simulated. The model generates steady growth or limit cycle behavior, depending on the assumptions made about elasticities of substitution and the saving rate.

WILLIAM S. MOORE II, Ph.D. Ohio State 1971. A general linear programming model of the manufacturing firm.

ROBINDRA N. MUKHERJEE, Ph.D. Rochester 1972. Optimal consumption and portfolio choices with transaction cost.

MOSES O. ODARO, Ph.D. Ottawa 1971. Intertemporal optimal consumption-saving allocation under uncertainty: An economic application of stochastic control theory.

Existing literature relating to types, sources, and measurement of uncertainty and their respective impact on intertemporal consumption-saving allocation is extensively surveyed. The problem is reformulated in terms of Sworder's maximum principle which yields a stochastic counterpart to the Keynes-Ramsey rule and allows the adjoint function to be interpreted as a marginal indirect utility function. Computational difficulties encountered in solving Sworder's stochastic partial differential equation require conversion of the original problem into a State Regulator framework.

MICHIHIRO OHYAMA, Ph.D. Rochester 1972. Stability and welfare in general equilibrium.

JOSEPH M. OSTROY, Ph.D. Northwestern 1970. Exchange as an economic activity.

To introduce trade into the theory of exchange, a decentralized general equilibrium model of the trading process is constructed. The point of departure is a mythical state of affairs in which the economy devotes no resources to organized markets. Traders have the opportunity to engage in a limited sequence of bilateral exchanges and are required to make their current decisions in ignorance of the tastes and supplies of their future partners. As a result, knowledge of market-clearing prices is not sufficient to guarantee achievement of market-clearing trades. Introducing a bookkeeping device into the model's trading arrangement, the store of value and means of payment function are demonstrated; money serves as a transferable signalling device through which the value of one's contributions can be recorded.

ANTHONY PAPPAS, Ph.D. Yale 1971. Adaptive behavior and speculative price formation.

WILFRIED M. A. PAUWELS, Ph.D. Columbia 1972. On the stability and efficiency of decentralized economic policy-systems.

Part I deals with the assignment problem, and examines the possibility of attaining a prespecified set of targets by assigning to each target variable a particular instrument variable which is then manipulated

with reference only to its assigned target variable. Part II examines, with the use of differential game theory, the performance characteristics of a policy system in which there are several autonomous policy-making institutions, each of them controlling some instrument variables, and each of them manipulating these instruments according to that institution's own preferences regarding the desirable behavior of the economy. The possibility of coordinating the decentralized policies in an uncertain environment is also examined.

URI M. POSSEN, Ph.D. Yale 1971. A synthesis of income-expenditure theory and neoclassical growth theory.

JACK J. PURDUM, Ph.D. Ohio State 1972. Investment in land enclosures: A study of five Nottinghamshire manors, 1783-1807.

CLARENCE G. RAY, Ph.D. South Carolina 1972. A human capital study of indentured servants in the American Colonies: 1607-1775.

This dissertation is a human capital study of the indentured servant in colonial America, with particular emphasis placed on Pennsylvania, Maryland, and Virginia. It is a comparison of free labor with that of bonded servants and is accomplished with the use of a cost-benefit equation. Even with a mortality factor of 50 percent, the benefits exceeded the costs when indentured labor was used instead of free labor.

CLYDE G. REED, Ph.D. Washington 1971. Price data and European economic history: England, 1300-1600.

JOHN W. REIFEL, Ph.D. Michigan State 1972. The demand for "free" goods.

YUNG WHEE RHEE, Ph.D. Johns Hopkins 1972. Absorptive capacity and optimum growth.

DONALD J. ROBERTS, Ph.D. Minnesota 1972. Lindahl equilibrium and the allocation of public goods with a measure space of consumers.

The chief result obtained is an existence theorem for Lindahl equilibrium. The approach to the proof involves the use of the dualities between the "market-clearing" and Pareto optimality conditions with public goods and between the Walrasian and Marshallian models of market adjustment. This approach, which relates to some recently studied adjustment mechanisms for public goods economies, permits a finite dimensional treatment, even though public goods prices may be points in an infinite-dimensional space.

HUGH T. ROCKOFF, Ph.D. Chicago 1972. The free banking era: A reexamination.

This study is concerned with the banking laws adopted in many states in the antebellum period that provided for free entry into banking and a bank currency secured by government bonds. An explanation of

the sudden inflations they occasionally produced is given. In addition, attempts are made to gauge their impact on the safety of the currency, the demand for money, the market for government bonds, and the allocation of bank capital.

STEVEN S. ROSEFELDE, Ph.D. Harvard 1972. Factor proportions and the commodity structure of Soviet international trade: 1955-68.

Factor proportion statistics generated in Soviet trade with the world, and other geographic entities are computed using the input-output procedures laid out by Leontief in his classic study of American factor proportions. This study demonstrates that the Soviet Union's factor proportions conform with the predictions of Heckscher-Ohlin theory; that is, the *USSR* exports relatively capital intensive goods to its capital poor trading partners, and labor intensive goods to capital rich nations.

JOEL B. ROSENBERG, Ph.D. Brown 1972. Research and advertising as stocks of knowledge and goodwill.

Four models are developed depending on whether knowledge and goodwill are stocks or flows. It is assumed that the firm is trying to maximize a discounted stream of profits over an infinite horizon. From these models, it is possible to ascertain the nature of these variables on the basis of observed statistical relationships. Empirical testing gave strong indications that both variables are stocks.

GARY R. SAXONHOUSE, Ph.D. Yale 1971. Productivity change in the Japanese cotton-spinning industry, 1891-1935.

DAVID SCHEFFMAN, Ph.D. Massachusetts Institute of Technology. Two essays in economic theory.

STEPHEN L. SHAPIRO, Ph.D. South Carolina 1972. The growth of the cotton textile industry in South Carolina: 1919-30.

The cotton textile industry continued to grow in South Carolina during the period 1919-30 in spite of a general decline in the industry. The continued growth in South Carolina was due to the abundant supply of white labor; labor that was willing to work for a low wage. The average wage differential between 1919-30 was 39 percent in South Carolina's favor. The supply of white labor was perfectly elastic at the going wage.

JOHN SHILLING, Ph.D. Massachusetts Institute of Technology. Investigations into a two-sector growth model with nonmalleable capital and two consumption goods.

HEATHER J. SLEMMER, Ph.D. Texas A&M 1971. The Cambridge criticism of neoclassical theory.

This dissertation investigates the criticism of those in Cambridge, England, to neoclassical theory as currently taught in the United States. The Cambridge criticism

challenges neoclassical theory on the grounds that the interest rate must enter the valuation of heterogeneous capital goods, and this may cause the wage interest ratio to be inversely related to the capital-labor ratio over some ranges of the production function. Therefore, as the capital-labor ratio increases, the permanently sustainable income-stream may decrease.

JAY B. SPECTOR, Ph.D. Michigan State 1971. A reformulation of Harrod growth theory.

A. MICHAEL SPENCE, Ph.D. Harvard 1972. Market signalling.

The essay develops an equilibrium model of signalling and information flows in job markets and examines the properties of these equilibria. Insights are applied to other markets with comparable informational properties. It is shown that multiple signalling equilibria exist, some are Pareto inferior to others, people tend to over-invest in things like education, the private return to signals exceeds their social productivity, and informational equilibria are potentially discriminatory in several qualitatively distinct ways.

ROSS M. STARR, Ph.D. Stanford 1972. General equilibrium analysis of monetary economies.

This study examines three topics: General equilibrium analysis of money eliminating need for double coincidence of wants in trade (*Quart. J. Econ.*, May 1972); Equilibrium and demand for media of exchange in an economy with transactions costs where trade may be chosen bilateral or multilateral; Equilibrium price of money in an economy with taxation. There are equilibria where the price of money is zero (money has no purchasing power). Taxes (payable in money) can assure positivity.

DALE E. SWAN, Ph.D. North Carolina 1972. The structure and profitability of the antebellum rice industry: 1859.

RASHESH B. THAKKAR, Ph.D. Rochester 1972. Structure of a two-sector three-factor-economy: Production, trade, and growth.

RINALDO H. TOPOROVSKY, Ph.D. Columbia 1972. Household capital formation: A theory (The HIT) and its empirical implications.

This dissertation provides six main theoretical and empirical contributions: review of the logical foundations of traditional demand theory; development of a general theory, including the traditional as a special case, of household-asset demand, defined as derived from that of household-produced services; elaboration of the control, variational and dynamic-programming problems of the household; estimation of household stock series; structural, LIML, testing of the postulated hypotheses; and determination of the HIT as a rich and useful theory.

GABRIEL TORTELLA, Ph.D. Wisconsin (Madison) 1972. Banking, railroads, and industry in Spain, 1829-1874.

JONG SOUE YOU, Ph.D. State University of New York (Binghamton) 1972. Money in the theory of economic growth: An analysis of comparative-dynamic and optimality aspects of growth equilibrium.

The thesis shows that an increase in the rate of monetary expansion raises the equilibrium capital-labor ratio and lowers the equilibrium per capita real balances and that the optimal rate of growth of interest-bearing money is equal to the nominal interest rate, if interest payment is financed by costless creation of money, but equal to twice this rate if it is financed by the increased productivity of capital resulting from the increase in real money stock.

THOMAS M. ZEPP, Ph.D. Florida 1971. Agricultural labor in the American South, 1860-1870: An analysis of the elasticity of substitution and change in functional shares.

This study presents estimates of the substitutability between labor (primarily slaves) and other inputs in the antebellum agricultural South and an approximation of the change in the functional labor share before and after the Civil War. The 1860 CES estimate is between one and two and is significantly different from zero. The functional labor share did fall absolutely after the Civil War but is found to have increased relative to the non-labor share.

### **Economic Growth and Development; including Economic Planning Theory and Policy, Economic Fluctuations and Forecasting**

IFTIKHAR AHMED, Ph.D. Iowa State 1972. Unemployment and underemployment in Pakistan.

MARVIN S. ANDERSON, Ph.D. Cornell 1972. The planning and development of Brazilian agriculture: Some quantitative extensions.

Development goals and instruments for Brazilian agriculture are identified and an attempt is made to quantify the relationships between the instruments and the objectives. A survey of the Brazilian agricultural economy in the period 1950-68 is included. Agricultural credit appears to be one policy instrument that has enhanced crop production.

R. KEITH AUFHAUSER, Ph.D. Harvard 1972. Work and slavery: Profitability, discipline, and technology on Caribbean plantations.

Slave and free labor are shown to resemble one another on the grounds of the motivation to work, the profitability of their employment, the methods of discipline, and the incentive to innovate. An examination of

Caribbean plantations in the nineteenth century provides the bulk of the data. The technology of production rather than the legal status of the worker is shown to determine the liberty of labor.

JAWAID AZFAR, Ph.D. Harvard 1972. The income distribution in Pakistan before and after taxes: 1966-67.

The dissertation analyzes the concentration of incomes in Pakistan in relation to theories on the relationship of growth and distribution, other estimates of income concentration in Pakistan and other developing countries. The incidence of taxes and their redistributive impact between income groups and regions is also studied.

DONALD A. BALL, Ph.D. Florida 1971. The economic impact of the American retirees in Jalisco, Mexico on the Mexican economy.

A personal field survey was made of a representative sample of American retirees living in the state of Jalisco, Mexico constituting approximately 50 percent of all American retirees in Mexico. An analysis was made of the income, expenditures, reasons for choosing Mexico for retirement, and other personal characteristics. From the sample taken, the total annual expenditures were estimated to be \$53 million. By use of multipliers, the total economic impact was computed to be \$376 million annually.

JOHN J. BERNARDO, Ph.D. Purdue 1972. Monopolists' intertemporal optimization with downwardly rigid money wages and aggregate price-employment relationships.

By eliminating the Walrasian postulate of complete information, a discrete-intertemporal labor market model is developed at the micro-economic level. The demand for labor is derived from a dynamic planning model for a monopolistic firm. Aggregation over firms yields a Phillips curve. The static and dynamic behavior of the Phillips relationship and of the individual decisions are analyzed with respect to policy changes and exogenous changes in the model's market and cost parameters.

CHARLES R. BLITZER, Ph.D. Stanford 1971. A perspective planning model for Turkey: 1969-84.

A dynamic multi-sector multi-skill planning model is developed for the Turkish case. Human capital formation is handled through endogenous schooling activities which are analogous with the investment activities for physical capital. The model chooses yearly levels of sectoral investments, human capital formation, export promotion, and import substitution. Since the model is optimizing, shadow prices are computed for all resources. Experiments are reported for parametric variation of the constraints and social welfare function.

AKE G. BLOMQUIST, Ph.D. Princeton 1971. Monetary policy and development financing: The case of Nigeria, 1950-66.

The study examines Nigeria's monetary and financial system and analyzes objectives and techniques of monetary policy in an open, low-income economy. Demand for money is studied empirically and the case for large-scale inflationary financing is considered quantitatively and found weak. The rules versus discretion problem is analyzed, and adoption of a modified, open-economy version of Friedman's proposed rule is recommended. Ways of increasing competition and efficiency in credit allocation are discussed.

RAJENDRA P. BODEPUDI, Ph.D. Wayne State 1972. The impact of the Mahalanobis two-sector planning model on labor reallocation in India: 1950-69.

LAWRENCE J. BRAINARD, Ph.D. Chicago 1971. Policy cycles in a planned economy: The case of Czechoslovak agricultural policy, 1948-67.

The study analyzes cyclical patterns in agricultural policy in Czechoslovakia. A theory of policy cycles in centrally administered economies is developed which highlights the interaction of political and economic variables in producing policy cycles. Four testable hypotheses are then derived from the theory. These hypotheses are tested in an analysis of the development of Czechoslovak agriculture in the postwar period.

J. JAMES BUCKNALL, Ph.D. Wisconsin (Madison) 1971. An appraisal of some of the developmental impacts the Kenya national trading corporation.

JOHN F. CHIZMAR, Ph.D. Boston College 1972. An econometric examination of organizational change in production functions of major industries: The Indian case, 1948-61.

BALTAZARA COLÓN-ZALDUONDO, Ph.D. Rutgers 1971. The saving and consumption variables in the growing Puerto Rican economy.

The study attempts to show how the consumption variable has to a large extent shaped the pattern of industrial development followed in Puerto Rico. The positive as well as the negative aspects of high consumption levels are examined. The problem of accelerating the rate of capital formation is viewed as one to be approached from the supply side (spectrum of feasible investment projects) rather than from the demand side (available investable funds).

LUIS R. CUCA-TOLOSA, Ph.D. Princeton 1972. Demography of Brazil: A regional study.

An estimate of mortality, natality, and internal migration for each state in Brazil and the country as a whole using United Nations Manual IV techniques and traditional methods of demographic analysis. Results show that quality of data is directly related with economic and social development and that mortality and natality are lower in more advanced states while migration goes from backward to advanced or less dense states.

FARHAD DAFTARY, Ph.D. California (Berkeley) 1971. Economic development and planning in Iran, 1955-67.

DIPANKAR DASGUPTA, Ph.D. Rochester 1972. On optimal economic growth.

PAUL R. DEUSTER, Ph.D. Wisconsin (Madison) 1971. Rural consequences of Indonesian inflation: A case study of the Jogjakarta region.

OKON J. ESSIEN, Ph.D. New York 1972. An econometric study of Nigeria's foreign trade, 1950-66.

The overall purpose of this study is to explore the structural relationship of the foreign sector of the Nigerian economy in order to assess the impact economic development and growth had in the period 1950-1966 on the country's trade relationships with the outside world. This calls for the determination of the relevant variables in the model and the ways in which they are related.

EFTHIMIOS C. FARANTZOS, Ph.D. Saint Louis 1971. Economic development and integration: The case of Greece.

This study is concerned with the impact of the association of Greece with the EEC—including welfare implications—from 1961 to 1966 on the Greek trade (imports) from the EEC countries and from the rest of the world. It was found that the association had an effect of very small economic significance. Also the relative welfare gains of superior resource allocation for Greece are minute and of very small economic significance.

DONALD S. FERGUSON, Ph.D. Cornell 1971. An economic appraisal of tick borne disease control in tropical Africa: The case of Uganda.

The complexities of identifying, planning, implementing, and evaluating a livestock development program in tropical Africa are examined. The study focuses on a project to control East Coast fever in Uganda, "the single greatest bottleneck to development" in that country.

ADOLFO A. FIGUEROA AREVALO, Ph.D. Vanderbilt 1972. Income distribution, employment, and development: The case of Peru.

This study deals with the problem of urban unemployment and underemployment in less developed countries. Specifically, its purpose is to investigate the effect of an income redistribution policy upon the demand for labor in the industrial sector of the Peruvian economy.

RICARDO FFRENCH-DAVIS, Ph.D. Chicago 1971. Economic policies and stabilization programs: Chile, 1952-69.

The study deals with Chile's economic policies from 1952 through 1969, a period covering three administrations and including three price-level stabilization pro-

grams. The study includes: a) a detailed view of the main features, goals and achievements, of each program; b) a description and analysis of economic policies (foreign trade, monetary, fiscal, tax, and income policies); and c) appendices where the econometric work and efforts made to homogenize and to build economically useful series are summarized.

CARLOS O. FONCK, Ph.D. Cornell 1972. Modernity and public policies in the context of the peasant sector: Honduras as a case study.

Effects of current development strategy on the subsistence farming sector in Honduras are examined through an analysis of peasant economic behavior at the farm level. Analysis focuses on short-run food grain sales response, which was found directly associated with output levels, unresponsive to price changes, and inversely related to price uncertainty. Agricultural policies in Honduras fundamentally aim at expanding farm output, but channels and means involved will not aid peasants in expanding their output.

HERBERT H. FULLERTON, Ph.D. Iowa State 1971. An economic simulation model for development and resource planning.

This dissertation was addressed to the problem of generating consistent socioeconomic information for economic development and resource planning. Special concern was focussed on the problem of information context, including spatial and temporal dimensions of this context. A modular type simulation model was developed and alternative futures for a state and four subarea sets were generated. The model features the flexibility of spatial and temporal context; the influence of area unique differences in competitive position on the pattern of change in output variables; the relationship between population change and labor requirement.

THOMAS G. GANIATSOS, Ph.D. California (Berkeley) 1971. Foreign-owned enterprise in Greek manufacturing.

LARRY K. GILBREATH, Ph.D. Florida 1971. Economic diversification on the Navajo Indian Reservation: A study of the development of small business enterprises.

This study analyzed the problems of small business development on the Navajo Indian Reservation. First, the reservation business community was statistically analyzed and then the data were compared with similar data from nearby counties, states, and the nation as a whole. The study then tested the hypothesis that the lack of small business development on the Navajo Reservation was largely the result of the Navajos' unique legal, financial, cultural, and educational institutions.

MORRIS GOLDSTEIN, Ph.D. New York 1971. A selective key-industry approach to anti-inflationary policy in the United States.

This dissertation explores a selective key-industry approach to anti-inflationary policy as one possible method of improving the short-run tradeoff between price stability and unemployment in the United States. The essence of this approach is that policy makers exert some control over the composition, as well as the level of aggregate demand when framing an anti-inflationary policy.

PETER M. GREENSTON, Ph.D. Minnesota 1972. The Food for Peace Program and Brazil: Valuation and effects of the commodity inflow.

The research focuses on measuring the benefits to Brazil over the period 1956-67 of the P.L.480 inflow within the context of the economic policies pursued by the Brazilian government. The first aspect involves the measurement of the pure aid component (i.e., unrequited aid) for the commodity flow. The second aspect is concerned with measuring the impact of the inflow upon domestic producer and consumer prices and thence upon quantity levels.

DAVID E. HANSEN, Ph.D. Iowa State 1971. Identification of economic impediments to utilization of Mexico's idle lands and policy implications for the Ejido and private sectors.

Regression results at state and county levels and case study interview findings support census reports that 40 percent of Mexico's cultivable lands are left idle. Reasons for idling land differ with land tenure class and include ownership uncertainty and extensive holdings in the private sector, and title uncertainty, labor and capital shortages, and attitudes unconducive to continuous cultivation in the Ejido sector. Changes are recommended in land tenure regulations and uses of traditional and nontraditional inputs.

JAMES J. HARRIS, Ph.D. Iowa 1972. Development of the Brazilian capital market.

The study examines Brazilian efforts to overcome deficiencies in the capital market and assesses the exportability of Brazilian methods to other developing countries. The central concern is the impact on Brazil's financial structure of major legislation during the post-1964 period. The legislation at issue—particularly Law 4728 and Decree Law 157—is examined within the context of the overall Brazilian environment and in terms of specific criteria developed for evaluation purposes.

OLI HAWRYLYSHYN, Ph.D. Massachusetts Institute of Technology. Patterns and determinants of internal migration in Yugoslavia.

PETER S. HELLER, Ph.D. Harvard 1972. The dynamics of project expenditures and the planning process: With reference to Kenya.

A problem often confronted in LDCs is an inadequate fiscal capacity to finance the recurrent costs of public sector projects. This is engendered by a "fiscal myopia," at the point of a project's inception, of the dynamic expenditure implications of public projects. This disser-

tation attempts to explain this phenomenon and uses Kenya as a case study, in a model of public sector growth which includes this induced expenditure relationship. This is empirically applied to the Kenyan Development Plan and the Kenyan Health Sector Plan.

BENTTI O. HOISKA, Ph.D. Columbia 1972. Technological change and the turnpike theorem.

This dissertation studies the impact of technological change on optimal growth in two-sector linear growth models. Technological change is incorporated into the model by introducing a new activity during the planning period. Optimality is determined with regard to terminal capital stocks and solutions are obtained by applying the maximum principle. A modified turnpike theorem is presented along with a discussion of how planners anticipate the new activity before it actually becomes available.

YOSHIHIDE ISHIYAMA, Ph.D. Stanford 1971. Social capital and economic growth.

AHMAD K. KATANANI, Ph.D. Iowa State 1971. Policies and models for planning the economic development of the non-oil sector in Saudi Arabia.

The non-oil sector in Saudi Arabia, although employing 46 percent of the labor force and supporting one-half to two-thirds of the population, contributes less than 10 percent to GNP and lags far behind the oil sector in development. Obstacles to development were identified and economic policies to help eliminate these obstacles were recommended. For a systematic approach to economic development a general interregional linear programming model was formulated. Modifications of the model were formulated to incorporate the top priorities in the Five-Year Plan.

KENNETH C. KEHRER, Ph.D. Yale 1972. Patterns of curriculum change during economic development.

MICHAEL I. KOLAWOLE, Ph.D. Cornell 1972. An economic study of tractor contracting operations in Western Nigeria.

Since the 1950's, attempts have been made to introduce the use of tractors into Western Nigeria. This study examined: the peasant farmer economic base and ability to use hire tractor service; the utilization of available power and factors that affect usage; the impact of tractor use on farm size, labor requirements, and farm practices; how timely farmers are serviced. A sample of 193 farmers and nine private tractor owners were interviewed. The major conclusions of this study were that overall performance and effectiveness have been below expectations; and that machine technology had been extended to peasant farmers who are not economically viable enough to absorb it.

MUNTASIR M. LABBAN, Ph.D. State University of New York (Binghamton) 1972. A development planning model for an underdeveloped economy with special reference to Lebanon.

The study analyzes the features, problems, and consequences of the service-oriented economy of Lebanon. It is argued that industrialization will help solve the unemployment, growth, distribution (per capita and regionally), and instability problems. The study advocates more active government intervention in the economy.

BARRY R. LAWSON, Ph.D. Cornell 1971. A stochastic model of settlement on the urban fringe: The role of land, its use, and related policy.

This study reviews the models put forward by others as a means of relating urbanization to the characteristics of alternative sites and their natural features in particular. An alternative, superior in many respects, is developed on the Markov Chain concept and an example, using data available for a Yugoslav urban region, is developed to show its usefulness in various policy making settings. Ideal data requirements do not appear difficult.

PO-CHIH LEE, Ph.D. State University of New York (Binghamton) 1972. Foreign assistance and economic development: The case of Taiwan.

This study presents the results of an econometric investigation of the development-through-aid of Taiwan in the period 1953-67. Its purpose is to assess the contribution made by foreign economic assistance to Taiwan's economic development and growth; and to provide, in the light of the Chenery two-gap theory, a more complete model for economic planning in aid-receiving developing countries.

STUART R. LYNN, Ph.D. North Carolina 1971. Aid requirements for development: The case of India.

The study analyzes "two-gap" theory and some models developed from it; applies the theory to an original model to determine foreign assistance requirements for India for the period 1969-79. Determinants of Indian saving, investment, imports, and exports are investigated, and regression equations offered for various categories of each. Alternative growth assumptions are formulated and applied to these regressions to project values. Results are unrealistic, but explained by the inability of the regressions to reflect recent trends.

ARTHUR J. MANN, Ph.D. Florida 1971. The income redistributional effects of the Puerto Rican fiscal system.

The purpose is to statistically estimate the income redistributional effects of the Puerto Rico fiscal (budgetary) structure. Tax, public expenditure, and net fiscal incidence by family "broad" income class are calculated. Tax and expenditure exportation is assumed in specific cases. The general pattern of net fiscal incidence is found to be regressive; i.e., the Commonwealth's fiscal system is so structured as to effectuate substantial income redistribution from lower to higher family income classes.

MICHAEL P. MAZUR, Ph.D. Massachusetts Institute of Technology. The economic development of Jordan.

O. ASTRA MEESOOK, Ph.D. California (Berkeley) 1972. Income distribution in Brazil.

FRANCIS W. MILLERD, Ph.D. Cornell 1972. An evaluation system for the natural resource sectors of the Prince Edward Island development plan.

Program planning and budgeting systems have only recently been considered for local governments. In the setting of a depressed economy and a massive regional development program, such a system is designed for the natural resources oriented programs of the Province of Prince Edward Island, Canada. While the system is based upon concepts of regional economic development, it is given a subregionalization format to fit the decision-making characteristics of the system in which it is used. Various problems in maintaining data validity and agency-treasury board (budget office) relations are considered.

TERRY D. MONSON, Ph.D. Minnesota 1972. Migration, experience-generated learning and infant industries: A case study of Turkey.

This study is an empirical examination of industrial learning patterns of semi-skilled Turkish workers employed in West Germany and in Turkey. From the analysis, it was concluded that experience-generated learning was more systematic; provided larger productivity increases; required shorter learning periods; and was less costly in Germany than in Turkey.

JOHN D. MORGAN, Ph.D. Georgetown 1972. A model for estimating centrally managed logistics costs to support alternative structures of Air Force operation forces.

BERNARDUS S. MULJANA, Ph.D. Iowa State 1972. The role of agricultural exports in Indonesia's economic development.

In the absence of export markets, Indonesia's agricultural export commodities would have zero or negative marginal value. Under the assumption that government consumption, nonagricultural exports, imports and capital inflows will take place according to the 1970-74 five-year plan, Indonesia's gross domestic product at 1968 prices may be expected to be around Rp 2,812.65 billion, or Rp 2,904.78 billion or Rp 3,074.64 billion, depending on whether agricultural exports will be US \$612.8 million, US \$675.1 million or US \$743 million.

KARIM A. NASHASHIBI, Ph.D. California (Berkeley) 1971. Foreign trade and productive efficiency under development planning: The case of the United Arab Republic.

ARTHUR C. NICHOLS, Ph.D. Georgia State 1971. An appraisal of the forecasting performance of the Council of Economic Advisers.

The objective of this study was to assess the accuracy of the annual forecasts of aggregate economic activity

in the United States prepared by the Council of Economic Advisers. The forecasts investigated were taken or inferred from the issues of the Economic Report of the President for the years 1962 through 1969; these forecasts included predictions of gross national product and its major components, the general price level, and the unemployment rate.

BADE ONIMODE, Ph.D. Ohio State 1972. Education, manpower, and the economic development of Nigeria, 1950-70.

GHATEKHA I. OSAYIMWESE, Ph.D. Northwestern 1971. A transportation-distribution problem: An application to the groundnut industry in Nigeria.

In this study linear programming is applied to the transportation problem of the groundnut marketing board in Nigeria. The problem was studied in the context of a firm, and thus some of the conceptual and measurement problems inherent in a macro-economic approach were avoided. The empirical analysis involved 1962 and 1970.

CHITLOYE A. OYEJIDE, Ph.D. Princeton 1972. Tariff structure and industrialization in Nigeria, 1957-67.

This study uses multiple regression analysis to evaluate the quantitative impact of tariff protection on Nigeria's industrialization process between 1957 and 1967. Indices used to measure industrial development include import-substitution ratio and growth rate of domestic production. Net effective protection rate measures tariff protection and two proxy variables measure market size. The regression results indicate that tariff protection accounts for 25 to 33 percent of the variations in the indices over the 1957-67 period.

GIRARD PESSIS, Ph.D. California (Berkeley) 1972. Limited control planning in a dynamic mixed economy.

F. MALIO H. PETRIE, Ph.D. Chicago 1971. Rates of return to physical capital in manufacturing industries in Argentina.

RAYMOND PRINCE III, Ph.D. North Carolina 1971. The effect of government policies to attract private investment into the manufacturing sector of Mexico, 1940-65.

The study attempts to determine the impact of Mexican government policies on net capital formation, improved technology, and greater labor productivity. Policies included in the model fall into three broad categories: subsidies; credit controls; and public investment. Dependent variables in the model represent either current private investment or changes in the efficiency index of some subsector with independent variables including changes in income-sensitive variables, output in other sectors, and governmental programs.

THOMAS G. RAWSKI, Ph.D. Harvard 1972. The economics of Chinese machine building, 1931-67.

This dissertation investigates the growth of China's machine-building industry between 1931 and 1967, with particular emphasis on the two decades following the creation of the People's Republic of China. The following topics are included: volume and composition of output during 1931-67; growth of capital stock and factor productivity during 1951-65; impact of excessive expansion of machine-building under the First Five-Year Plan (1953-57); effect of institutional factors on enterprise activity; and quality and assortment of machinery production.

COSMAS L. ROBLESS, Ph.D. Indiana 1972. Malaysian development alternatives, 1971-80.

The object of the dissertation is to evaluate a number of alternative growth paths open to Malaysia during 1971-80 and the general policy implications of each of the options. The core model employed was the two-gap planning model, the two gaps being the investment-savings and import-export gaps.

STEPHEN R. SACKS, Ph.D. California (Berkeley) 1971. Entry of new competitors in Yugoslav market socialism.

SAKDA SAIBUA, Ph.D. Wisconsin (Madison) 1972. Optimal patterns of growth and external indebtedness: The case of Thailand.

MARJORY E. SEARING, Ph.D. Georgetown 1972. Education and its contribution to economic growth under socialism: The experience in Hungary and Poland.

DON M. SHAKOW, Ph.D. California (Berkeley) 1972. Optimal transformation of welfare criteria: A normative analysis of pricing under NEP.

FAROUK M. SHALABY, Ph.D. Oklahoma State 1972. An evaluation of the contribution of United States Public Law 480 to the food grain trade consumption, and production of the less developed countries.

The objectives of this study are to determine whether or not P.L. 480 shipments to the less developed countries have substituted or supplemented commercial grain imports; to estimate the contribution of concessional sales under the law to the LDC's food grain consumption; and to analyze, in light of the P.L. 480 trade and consumption effects, where, how, and under what conditions P.L. 480 might have influenced domestic food grain production in these countries.

LARRY R. SHOTWELL, Ph.D. Ohio State 1971. The defense spending process.

INDERJIT SINGH, Ph.D. Wisconsin (Madison) 1971. A recursive programming model of traditional agriculture in transition: A case study of Punjab, India.

LESLIE E. SMALL, Ph.D. Cornell 1972. An economic evaluation of water control in the northern region of the Greater Chao Phya Project of Thailand.

Information derived from farm surveys and government reports was utilized to estimate the rate of return to past investments in water control facilities in the northern part of the central plain region of Thailand. In addition, alternative strategies for further development of the water control system were evaluated from the standpoint of their effect on the rate of adoption of improved production practices and double-cropping.

DONALD SNYDER, Ph.D. Pennsylvania State 1971. An econometric analysis of household consumption and saving in Sierra Leone.

SPENCER STAR, Ph.D. California (Berkeley) 1971. The sources of growth in U.S. manufacturing industries, 1950 to 1960.

WAYNE R. THIRSK, Ph.D. Yale 1972. The economics of Colombian farm mechanization.

JOHN E. TODD, Ph.D. Yale 1972. Efficiency and plant size in Colombian manufacturing.

KHAIRY A. TOURK, Ph.D. California (Berkeley) 1971. A new pricing system for the Suez Canal.

KUC-CHENG TSENG, Ph.D. Pennsylvania State 1971. An analysis of the growth of selected export industries in Taiwan, 1952-69.

DAVID J. VAIL, Ph.D. Yale 1971. The public sector as stimulus of innovation adoption in African small-holder agricultures: A case study of Teso District, Uganda.

JOSEPH L. C. VELLIN, Ph.D. Cornell 1971. A full employment strategy for agricultural development in Mauritius.

This research uses Mauritius, an extreme case of tropical monoculture, to test the hypothesis that problems of export earnings and employment can be ameliorated through an agricultural development strategy involving both import substitution and export specialization. A linear programming model is used to analyze simultaneously the effect on employment and foreign exchange earnings of alternative land use patterns. The results, obviously oversimplifications of reality, were positive.

ASIM YUCEL, Ph.D. New York 1972. An analysis of Turkish industrial progress: The problem of financing and the role of monetary and credit policy, 1923-67.

The objective of this study is to analyze Turkish industrial progress during the 1923-67 period concentrating on an investigation of how the government's monetary and credit policy affected industrial development by providing financial resources to industry and expanding aggregate demand through extension of bank credits.

RAUL YVER, Ph.D. Chicago 1971. The investment

behavior and the supply response of the cattle industry in Argentina.

### Economic Statistics; including Econometric Methods, Economic and Social Accounting

WILLIAM J. BARGER, Ph.D. Harvard 1972. The measurement of labor input: *U.S. manufacturing industries, 1948-66*.

The relationship between indexing theory and dynamic neoclassical models of producer behavior is investigated. A theoretically consistent index structure is developed and applied to the measurement of labor input in *U.S. manufacturing*. A measure of labor quality is developed, and postwar changes in labor quality are analyzed as a function of the changing sex-age-education distribution of the labor force. A comprehensive set of labor quantity and price indices is also presented.

CHARLES M. BEACH, Ph.D. Princeton 1972. Estimating distributional impacts of macro-economic activity.

An indirect quantile approach is used to estimate impacts upon the structure of income concentration of changes in aggregate unemployment and participation rates and average wage levels. A model of the channels through which fluctuations in these aggregates affect individual quantiles is estimated for males disaggregated by age over 1947-70, and conclusions about the implied systematic short-run behavior of income concentration are then derived from the estimated quantile equations.

DAVID F. BURGESS, Ph.D. Wisconsin (Madison) 1972.

An analysis of the composition of nonhuman wealth owned by the United States private household sector in the postwar period.

JOHN J. FARRELL, Ph.D. Catholic 1972. The scope and functions of an educational price index: The Maryland experience.

SHAKIL A. FARUQI, Ph.D. Rutgers 1972. Life insurance portfolios: A study in non-linear estimation

Parametric nonlinearities generated by lag distributions and autoregressive processes have been treated in a concurrent fashion in various specifications of a portfolio model of *U.S. life insurance companies*. Scanned estimates of the model have been obtained for the period 1957-70 using an elaborate replication routine designed specifically for the purpose. The results emphasize inadequacies of the classical procedures if in a system we are faced with dual nonlinearities emanating from the postulated behavior of explanatory variables and stochastic errors.

DAVID K. FOOT, Ph.D. Harvard 1972. Interactive investment behavior: A study of *U.S. manufacturing industry*.

This study uses data for the postwar *U.S. manufacturing industry* to estimate an interactive investment

model by rejecting the traditional (long-run) assumption that replacement investment is proportional to the lagged capital stock. It combines a model that explains the timing of replacement and modernization investment with an expansion investment model and shows that the two decisions are interdependent. Detailed sensitivity analysis, simultaneous estimates, and elaborations on the basic two-equation model are presented, together with estimates for selected S.I.C. two- and three-digit level industries.

B. R. FROELICH, Ph.D. Minnesota 1971. Estimation of a random coefficient regression model.

The advantages of specifying a random coefficient (rather than the usual fixed coefficient) regression model in econometric work is discussed. Two sets of unknown parameters need to be estimated: estimates of the mean response coefficients and an equal number of variance estimators. Several large sample properties of ordinary least squares estimation of such a model are given. A Monte Carlo experiment is designed to obtain information of various estimation procedures with small samples.

THEODORE GAMALETOS, Ph.D. Wisconsin (Madison) 1972. International comparison of consumer expenditure patterns: An econometric analysis.

WOO BONG LEE, Ph.D. Rutgers 1972. Competitive spatial equilibrium analysis for an agricultural subsector: Econometric estimation of regional demand and supply relations and application of non-linear quadratic programming.

Within the framework of location theory, pure trade theory, and production theory, the non-linear spatial equilibrium model and its programming problem are elucidated with an empirical application of the model to an agricultural subsector of the *U.S. economy*. In the study, regional demand and supply functions are estimated and problem of efficient allocation among spatially separated markets is solved by quadratic programming formulation to maximize the net sum of regional products.

AN-LOH LIN, Ph.D. Rochester 1972. Distribution, interpolation, and extrapolation of time-series by related series.

ALLAN I. MENDELOWITZ, Ph.D. Northwestern 1971. The measurement of economic depreciation in the United States.

This paper attempts to improve the national income and product accounts by computing economically meaningful estimates of depreciation. A model based on the generally accepted notion of depreciation as a change in value was developed. Data were the firm level observations of the McGraw-Hill Capital Expenditure Survey, and OLS estimation was used. Findings suggest that *U.S. tax depreciation* is too rapid and that national income account depreciation is overstated.

GORDON H. OTTO, Ph.D. North Carolina State 1971. Toward an endogenous lag theory in capital investment behavior.

Jorgenson's neoclassical model of capital adjustment has been restructured as a discrete-time mathematical programming problem. There are three significant results: the lag distributions which emerge are endogenous to the model and are functions of its parameters; the upward and downward adjustment paths are induced by different factors and hence are different, that is, the adjustment process is not symmetric; and third, both processes are of the lead-lag variety rather than the strictly lagging variety which is commonly employed in empirical analysis.

JON K. PECK, Ph.D. Yale 1972. Comparison of alternative estimators for a dynamic relationship estimated from a time-series of cross-sections when the disturbances are small.

MICHAEL PRELL, Ph.D. California (Berkeley) 1971. Relative movements of *U.S.* price indexes in the postwar period.

TEJBHAN S. SAINI, Ph.D. New School 1972. The estimation and evaluation of the Cobb-Douglas production function from inter-plant cross-section data: A case study of lumber and plywood industries.

This study computes inter-plant cross-section estimates of production functions specified in several alternative ways. Among these are experiments which define "output" in physical units, value-added units, and as "installed capacity." It confirms previous suspicions that such econometric estimates have major defects from the standpoint of summarizing the production possibilities of an industry. It also reveals the incorrectness of Douglas' work since his basic formulation is shown to militate against the accepted concepts of micro-economic theory.

TOM S. SALE III, Ph.D. Louisiana State 1972. A study of the size distribution of income: The case of Louisiana, 1950-70.

The purpose of the study is to explain the socioeconomic structure associated with intrastate differences in Louisiana's family income distribution and to analyze changes in concentration over the period 1950-70. Nine independent variables are tested in three cross-section analyses. Median family income and median years of education rank consistently high in each of the cross-sections but all variables show reduced significance in explaining rates of change in inequality.

WILLIAM B. STRONGE, Ph.D. Iowa State 1971. Macroeconomic simulation of Irish dependence on the United Kingdom.

A macro-econometric model of the British Isles was estimated to provide information on the impact of changes in *U.K.* conditions on the Irish economy. The model was of the simple Keynesian type with Irish

exports functionally related to *U.K.* imports. It was estimated using ordinary least squares on quarterly data. Irish quarterly national accounts were obtained by interpolation. The *ex post* forecasting ability was tested. Multipliers were obtained and a simulation procedure was estimated.

JACK J. TAWIL, Ph.D. California (Los Angeles). The small-sample behavior of various estimators in a linear structural and a linear functional relationship: A Monte Carlo study.

For seven major methods (comprising thirty-four estimators) for estimating the linear structural and linear functional relationship, namely, maximum likelihood, least squares, instrumental variables, analysis of variance components, grouping, product cumulants, and restrictions on the various distribution functions, estimates are calculated and compared for a variety of points in the parameter space. Confidence intervals are constructed and tested for nearly all of the estimators. Finally, two new estimators are derived, and they are found to have, under restrictive conditions, excellent small-sample properties.

WALTER J. WADYCKI, Ph.D. Northwestern 1971. Jan Tinbergen and early econometric testing: A reconsideration.

KENNETH R. WHITE, Ph.D. Oklahoma 1971. The effects of certain specification errors on the properties of parameter estimates in small samples of a single equation model.

The effects of two specification errors, each involving the error term, are analyzed to determine their combined effects on parameters and forecasting. The two specification errors are those of heteroscedasticity and a correlation between the independent variable and the error term. A single equation Monte Carlo analysis is used to determine the joint effects of the specification errors in a least squares analysis.

ROLAND Y. WU, Ph.D. Stanford 1971. The implication of linear aggregation on predictions from econometric models.

Two types of linear aggregation, in the context of regression analysis, are considered in this study: contemporary aggregation over  $N$  units and temporal aggregation over  $M$  nonoverlapping periods. Using the expectation of mean-squared error of prediction as a criterion, the study compares the efficiencies of predicting the aggregate dependent variable from aggregate and disaggregate models.

### Monetary and Fiscal Theory, Policy, and Institutions

EARL W. ADAMS, JR., Ph.D. Massachusetts Institute of Technology. "Reduced form" tests of monetary and fiscal actions: The case of Italy.

PAUL G. ALTHAUS, Ph.D. Duke 1972. Fiscal and monetary policy in economic growth.

ROBERT E. ANDERSON, Ph.D. Johns Hopkins 1972. The individual's transactions demand for money: A utility maximization approach.

MUKUL G. ASHER, Ph.D. Washington State 1972. The concept of fiscal leverage and its application to the fiscal policy performance of the government of India, 1950-51 to 1966-67.

The approach in this thesis is similar to Richard Musgrave's leverage analysis but it extends his work by providing a possible general equilibrium model, by adding a foreign sector to the model, and by applying the model to the Indian economy in both a year-to-year analysis and over the three five-year planning periods.

ROBERT N. BARONE, Ph.D. Georgetown 1972. An economic approach to stock price forecasting.

JAMES R. BARTH, Ph.D. Ohio State 1972. Cash adjustments of commercial banks.

JOSEPH R. BISIGNANO, Ph.D. Stanford 1971. The portfolio behavior of nonbank financial institutions.

This dissertation analyzes the financial asset portfolio behavior of several nonbank financial intermediaries satisfying recent criticisms of Tobin and Brainard. Static and dynamic interrelated adjustment models are derived. Sectoral aggregate and disaggregate models are estimated and simulated satisfying wealth and price constraints across equations for both estimated short-run and imputed long-run coefficients. Adjustment speeds are in sharp contrasts to those derived from single asset stock adjustment models. A separate model for life insurance commitment behavior is also derived and estimated.

HAROLD A. BLACK, Ph.D. Ohio State 1972. The money supply process in Great Britain: 1951-69.

JOHN A. BROADBUSH, JR., Ph.D. Indiana 1972. A stochastic model of individual bank behavior.

The purpose of the study is to develop a model of individual bank balance sheet behavior which explains all aspects of the bank's balance sheet decision process. The fundamental assumptions are that the bank acts to maximize the return to equity over a specified time period subject to the balance sheet identity constraint, and that the time period is long enough to permit control over each category of asset and liability. The model takes explicit account of revenue and cost flows associated with each asset and liability stock.

STEPHEN A. BUSER, Ph.D. Boston College 1972. Separation, diversification, and decomposition in the single-period portfolio problem.

The dissertation addresses two problems of single-

period portfolio theory: optimal allocation over a fixed number of securities; and the optimal number of securities. For each problem a single-step solution procedure is developed for a multiparameter model that generalizes the various mean/variance models. Efficiency rules, independent from the investor's subjective preferences, are established through separation theorems. Sufficient conditions are presented for the decomposition of the problem to allow divisional specialization.

SULEIMAN M. BUSHNAQ, Ph.D. Catholic 1972. Monetary versus industrial demands for gold.

LARRY B. BUTLER, Ph.D. California (Berkeley) 1972. A comparison of the inventory and asset theories of the demand for money.

JACK CARR, Ph.D. Chicago 1971. A dynamic monetary model of business fluctuations.

WARREN L. COATS, JR., Ph.D. Chicago 1972. The September 1968 changes in "Regulation D" and their implications for money supply control.

New reserve requirement calculation procedures (lagging by two weeks the demand deposits used in calculating required reserves and the vault cash eligible toward meeting it) have weakened Federal Reserve control of the money supply, increased the need for defensive open market operations, and increased the volatility of money market rates. The dissertation is a theoretical and empirical study of the implications of these new rules.

RICHARD V. L. COOPER, Ph.D. Chicago 1971. Money and stock returns.

The purpose of this thesis is to examine the relationship between the money supply and returns to equity and to explore the usefulness of spectral analysis in econometrics. The quantity theory of money is combined with the theory of efficient capital markets, which suggests that stock returns may lead changes in the money supply if changes in the money supply can be successfully anticipated. Cross-spectral results indicate that stock returns do lead changes in the money supply.

JOSEPH M. CREWS, Ph.D. North Carolina 1972. Alternative optimal open market strategies: A simulation approach.

The FRB-MIT model is transformed into optimizing made by a non-linear extension of Theil's certainty equivalence theorem. Alternative policy regimes (instrument proxies) and policy strategies (intermediate targets) are simulated over a 16-quarter period. Significantly different welfare levels are attained under alternative regimes and strategies. The Fed is found to respond to targets which, in its view, are important in the policy transmission mechanism, but could improve its performance by adopting a money supply strategy.

CHRISTOPHER CURRAN, Ph.D. Purdue 1972. *U.S. municipal expenditure and revenue patterns: 1900 to 1930.*

The financial data of all cities over 30,000 population is examined in detail for the first 30 years of this century. Special emphasis is placed on the correction and expansion of the data related to municipal finances of earlier scholars such as Fabricant and Goldsmith. The data also has been examined for effects related to the size and regional location of the cities. Finally, the data was examined for any indications of economies of scale in city operation.

BRUCE C. DIEFFENBACH, Ph.D. Harvard 1972. *Uncertainty and relative rates of return on securities: A theoretical and empirical analysis.*

A model of liquidity premiums on securities in a multiperiod market equilibrium with uncertainty was constructed. The formal reduction by dynamic programming of an individual's multiperiod consumption-investment decision to a single period decision provides the foundation for the analysis. Covariances of security returns with aspects of the state of the world determine liquidity premiums. Empirical testing and estimation of the model and an application of the theory to the term structure of interest rates conclude the dissertation.

LOUIS H. EDERINGTON, Ph.D. Washington (St. Louis) 1972. *Yield and underwriter spreads on new issues of corporate bonds.*

New issues during the period 1965 through 1969 were examined. Particular attention was paid to the effect of uncertainty regarding the demand for the new issue, and it was concluded that both spreads varied directly with the level of uncertainty. Other hypotheses were also tested including several relating to the impact of both spreads of inter-syndicate competition and the manner of offering.

SANFORD M. EDGAR, Ph.D. Oklahoma 1971. *A critique of Federal Reserve-Federal Open Market Committee activities in the post-Accord period: 1953-1969.*

The interpretation of "indicators" and "targets" of monetary policy is discussed to investigate why theoreticians do not agree on the appropriate economic variables for these roles. An econometric model of the financial sector, introducing dynamic open market operations and the discount rate as the primary policy tools, is used to analyze Fed operations. One conclusion was that the discount rate, not OMO, seems to be the primary tool for making dynamic quarterly changes in policy activity.

NORMAN H. ERB, Ph.D. Indiana 1972. *Change in U.S. banking structure, 1956-67.*

This dissertation examined changes in commercial banking structure brought about by entry of new firms and by *de novo* branching and merger in selected states. The purpose was to investigate the fate of unit banks in the United States, with special attention given to

Indiana, Missouri, North Carolina, New Hampshire, New York, and Virginia, where recent statutory changes had increased the degree of branching permissiveness. The inquiry tested the hypothesis that permitted extension of multi-office banking leads to a steady reduction in the number of small unit banks, and that the extension of branching permissiveness would lead to the ultimate end of unit banking in the United States.

RICHARD L. FLOYD, Ph.D. Iowa State 1972. *The demand for financial assets: A portfolio approach.*

The purpose of this study was to estimate the demand for various assets which compose the net non-human wealth of the private sector. A systematic attempt to develop a Smith-Tobin model of the portfolio was made and an expectations framework incorporated a distributed lag technique for estimation purposes. The Almon technique was adopted and results presented from two alternative specifications of the model.

RICHARD T. FROYEN, Ph.D. Maryland 1972. *A quantitative model of the money-income relationship.*

This study presents a model to provide a detailed quantitative description of the linkages between monetary and real variables which incorporates the views of the "monetarists." The implications of the model are that both monetary and fiscal policy have significant effects on income. However, neglect of the monetarist views will give an upward bias to the measurement of both monetary and fiscal policy multipliers. The bias is greatest in the case of monetary policy.

CLIFFORD L. FRY, Ph.D. Texas A&M 1972. *Domestic monetary and financial theory and institutions.*

This research is a theoretical and empirical examination of the currency-deposit ratio. In the theoretical portion the responses of the ratio to changes in determining variables are derived. In the empirical portion past responses of the ratio to key variables are determined and discussed. In addition, the explanatory powers of various measures of the holding cost of demand deposits are investigated.

ARTHUR E. GANDOLFI, Ph.D. Columbia 1972. *The demand for commercial bank deposits during the great contraction.*

In this study a single equation, least squares technique was used to estimate yearly demand functions for the 1929-33 period from cross-sectional data. We defined the demand for deposits to be a function of state permanent income, state population, the rate of interest paid on deposits, and the rate of bank failures. We concluded that the demand function during this period was stable and that the American economy was neither in nor approaching a liquidity trap situation. We also concluded that bank failures had a greater affect on time deposits than they did on demand deposits.

ROBERT J. GENETSKI, Ph.D. New York 1972. *An anal-*

ysis of the role of economic factors in commercial bank portfolio behavior.

The role of economic factors in the portfolio decision-making process of commercial banks is analyzed. The analysis explores the extent to which various bank asset and liability categories respond to fundamental economic variables, and the consistency of this response with respect to widely accepted economic theory. Furthermore, an effort is made to determine the effect of various degrees of aggregation upon the responsiveness and significance of changes in these categories.

PAUL C. GRIER, Ph.D. New York 1971. Large block trades and price behavior of NYSE listed stocks.

The securities markets with their almost continuous flow of information on prices and quantities provide the economist with one of the most fertile fields for the study of price determination processes. This dissertation will examine one aspect of security price determination—the effects of large blocks of shares on the continuity and temporal structure of prices.

ROBERT P. INMAN, Ph.D. Harvard 1972. Four essays on fiscal federalism.

Essay I explores the problems caused by autonomous local government decision making in a regional economy which allows fiscal base disparities. Essay II develops a model of local government fiscal behavior based on the optimization of a preference function subject to a budget constraint. The model extends the previous work by explicitly introducing adjustments in the city's own revenues as well as changes in public service outputs. Essays III and IV are econometric studies of local fiscal behavior based on a sample of the forty-one major U.S. cities for the 1966-67 budget period.

JERRY W. JOHNSON, Ph.D. Iowa State 1971. Further evidence as to the relative effects of monetary versus fiscal policy.

This study provides additional evidence in the debate over the relative effects of monetary versus fiscal policy. As fiscal policy indicators, changes in federal government expenditures on goods and services, changes in federal government tax receipts and transfers, approximated through the use of the tax and transfer functions of the Fed-MIT econometric model, were used. Four alternative monetary indicators were used in the derivation of the required reduced forms. The reduced forms were estimated using the Almon lag technique with the result that the role of money was more strongly supported than the role of the fiscal variables

STEPHEN E. KAGANN, Ph.D. New York 1972. A micro-theoretic analysis of the differential effects of tight money on the loan policies of commercial banks.

Employing Lagrangian techniques of maximization, with a variable-level demand function, it is shown that, in the face of tight money, banks with at least a two-period horizon will prefer to lend to large firms which can use their power to play one bank off against another

even though the marginal return on loans to small but powerless borrowers is significantly greater. A policy of fixed proportions should be implemented along with credit restriction.

LIONEL KALISH III, Ph.D. Washington (St. Louis)

1972. The influence of current and potential competition on a commercial bank's operating efficiency.

In this paper hypotheses are developed which relate the current and potential competition in a bank's market to that bank's operating efficiency. A frontier estimation procedure is used to develop an inefficiency index. Such a procedure also produces information about economies of scale and organizational form in banking. Current competition is measured by concentration ratios, and potential competition by state branch banking laws. Hypotheses are tested by multiple regression analysis with the inefficiency index as the dependent variable.

EDI KARNI, Ph.D. Chicago 1971. The value of time and the demand for money.

This dissertation studies the hypothesis that, *ceteris paribus*, the demand for real money holdings is positively related to the real value of time. This relationship reflects attempts of households and firms to save time in conducting transactions when the relative price of time increases. An inventory model of money demand which incorporates the value of time is tested against U.S. time-series and international cross-section data. Broadly speaking, the empirical evidence supports the hypothesis.

KENNETH KLEEFELD, Ph.D. Wisconsin (Madison)

1972. A study of the postwar demand for financial assets by households and institutional investors.

RAYMOND J. KRASNIEWSKI, Ph.D. Purdue 1972. The derivation and application of measures of revenue capacity and relative effort for local governments in Indiana.

By applying statewide average rates to each local financing source, this thesis developed estimates of local government revenue-raising ability. Effort measures were calculated as index numbers relating actual collections to estimated capacity. The fiscal data were then incorporated into three grant-in-aid formulas: federal revenue sharing, federal assistance to impacted school districts, and a state program for property tax relief.

TERRENCE A. LARSEN, Ph.D. Texas A&M 1971. Credit market efficiency and the allocation of resources: A qualitative approach.

A qualitative framework of a general economic system is constructed, incorporating financial intermediaries and emphasizing interrelationships between "real" and "financial" sectors. Demand and supply equations for all markets are constructed and solutions are in terms of the directions of change in the price level

and market rates of return with respect to exogenous shifts in the money supply, reserve levels, etc.

NORMAN B. LEFTON, Ph.D. Chicago 1972. The demand for real cash balances and the expected permanent and contemporaneous rates of inflation.

This study examines the manner in which the rate of inflation affects the demand for real money balances. It differs from previous investigations in that it distinguishes between long- and short-term expectations regarding inflation. Moreover, at both the empirical and theoretical level, it establishes a case for questioning the validity of the assumption of a negative coefficient for the expected rate of change of prices in the standard demand for money equation.

THOMAS J. LENGVEL, Ph.D. Louisiana State 1971. Federal Reserve "even keel" policy: An historical and empirical analysis.

MAURICE D. LEVI, Ph.D. Chicago 1972. Inventory disequilibrium and the effects of monetary and fiscal policy.

If monetary and fiscal policy affect spending, and firms meet changes in spending from stocks as well as production, then even if policy affects spending consistently, output could be variably affected. It is found that policy affects sales more consistently than output. Further, if production and inventory accumulation are smoothed, then if policy affects sales it should also affect inventories. It is found that effects of successful policy can be determined through inventory behavior and not via direct analysis on sales or output.

WILLIAM A. MCCOLLOUGH, Ph.D. Florida 1971. Florida banking: Present market structure and performance and an inquiry into the probable effects of alternative forms of bank organization.

A theoretical framework is developed which includes an optimizing condition. The objective function is specified and the constraints are enumerated. Published banking data provides the basis for establishing the characteristics associated with the uni-office banking of Florida and the multi-office alternative. These characteristics are then incorporated in the objective function. The results indicate that the multi-office form of bank organization would be more optimal in Florida than the current uni-office constraint.

FRANCIS L. McDONALD, Ph.D. Georgetown 1972. The term structure of interest rates.

MARSHALL E. MCMAHON, Ph.D. Vanderbilt 1972. Federal Reserve behavior, 1923-31.

This study attempts to increase understanding of the failure of the U.S. Federal Reserve System to undertake a more expansionary monetary policy after the economic downturn of 1929. While several contradictory explanations of this behavior have been offered, none of these has been subjected to a rigorous statistical test. In this thesis, a model of Federal Reserve behavior is formulated and tested against the available data.

MICHAEL K. MADDEN, Ph.D. Iowa State 1971. A portfolio explanation of the behavior of stock prices.

JEAN M. MALEY, Ph.D. Rochester 1972. The impact of federal grants on provincial budgets: Canada.

JOHN R. MALKO, Ph.D. Purdue 1972. Allocating municipal fire protection expenditures to business firms and households according to a benefits-received criterion.

GHANSHYAM MEHTA, Ph.D. California (Berkeley) 1971. The structure of the Keynesian revolution.

MARCUS H. MILLER, Ph.D. Yale 1971. An empirical analysis of monetary policy in the United Kingdom, 1954-65.

SHIGEO MINABE, Ph.D. State University of New York (Binghamton) 1972. A comparative analysis of Keynesian and Swedish theory of economic fluctuations.

The present study is a critical examination of the post-Keynesian and the Stockholm School of Economic Fluctuations. It is known that some Swedish economists initiated a Keynesian revolution even before Keynes himself. However, the contributions made by such people as Ohlin should not be compared with those of Keynes' in the light of the static criteria of the Keynesian revolution. Ohlin's analysis goes far beyond comparative statics toward post-Keynesian dynamics. Also in this study, we have improved the widely accepted post-Keynesian cyclical growth theory that was expounded by Duesenberry, Goodwin, and Matthews by examining the short-run dynamics of the savings function.

WARREN E. MOSKOWITZ, Ph.D. Massachusetts Institute of Technology. The theory of compensating balances.

ANTON S. NISSEN, JR., Ph.D. Princeton 1971. Government securities dealers and the term structure of interest rates.

The dissertation tests the hypothesis that the size and maturity composition of government securities dealers' holdings of direct Treasury issues are influenced by expected interest rates and that securities of differing maturity are substituted on the basis of such expectations. Equations are estimated for monthly average holdings in several maturity categories for the 1961-68 period under three naive assumptions concerning expectations. The results provide relatively little support for the hypothesis, but it is suggested that they may reflect inappropriate expectational assumptions rather than a true lack of the hypothesized response.

JAMES R. OSTAS, Ph.D. Indiana 1971. The relationship between the supply of mortgage credit and non-interest mortgage terms.

RICHARD E. PETERSON, Ph.D. California (Berkeley)

1972. The permanent-income hypothesis of the demand for money.

JOHN REA, Ph.D. Wisconsin (Madison) 1972. Factors contributing to cycles in the stock of money, 1951-69.

DON C. READING, Ph.D. Utah State 1972. A statistical analysis of New Deal economic programs in the forty-eight states, 1933-39.

This dissertation uses loan and expenditure data for each of the forty-eight states for all New Deal programs for the 1933-39 period. Cross-sectional analysis of the data suggests that, given the representative variables selected, New Deal agencies failed to direct expenditures and loans in a manner which effected reform, but did expend in a pattern that would contribute to relief and recovery and at the same time improve the utilization of natural resources.

GORDON S. ROBERTS, Ph.D. Boston College 1972. A general equilibrium theory of asset prices with applications to the theory of the term structure of interest rates.

LEWIS J. ROSEN, Ph.D. Yale 1971. Stock market capital gains and consumption expenditures.

CHARLES R. ROSS, Ph.D. Oklahoma State 1972. The effects of state and local government expenditures on the distribution of income in Oklahoma.

The purpose is to estimate the effects which state and local government expenditures have on the distribution of income in Oklahoma. The major part of the study is an attempt to establish criteria with which to distribute among income deciles the benefits from various categories of government expenditure. Theoretical explanations of what groups benefit from various types of government expenditures are determined and then estimates of how they are distributed among income deciles are derived.

DANIEL L. RUBINFELD, Ph.D. Massachusetts Institute of Technology. An econometric analysis of the market for general obligation municipal bonds.

GARY J. SANTONI, Ph.D. New Mexico 1972. The demand for money: A study of paradigms.

This dissertation sets out to establish the origin of the controversy that exists with respect to the channel(s) through which monetary phenomena influence economic activity. The origin is fixed in the 1930's and identified with the emergence of two schools of thought: Price Theoretic Behaviorism and Nominalism. During the course of the analysis it is argued that the traditional distinctions drawn between Fisherian quantity theory and Keynesian theory are essentially superficial.

IAN G. SHARPE, Ph.D. Stanford 1971. An econometric model of the Australian monetary sector with particular emphasis on the mortgage commitment process.

JEREMY J. SIEGEL, Ph.D. Massachusetts Institute of Technology. Stability of a monetary economy with inflationary expectations.

MYRON B. SLOVIN, Ph.D. Princeton 1972. Deposit rate setting at financial intermediaries: Theoretical models and econometric analyses with additional focus on savings and loan associations.

This dissertation is a theoretical and econometric analysis of the deposit rate setting behavior of financial intermediaries. Several theoretical models of rate setting are formulated and used as a basis for the estimation of deposit rate equations. Cross-section results are obtained for a sample of savings and loan associations, while time-series results include equations for commercial banks and mutual savings banks as well as savings and loan associations.

GARY N. SMITH, Ph.D. Yale 1971. Estimating a general disequilibrium model of the financial sector.

LEWIS J. SPELLMAN, Ph.D. Stanford 1971. Finance as an industry: A simple model of growth.

Finance is analyzed by the introduction of a financial sector into a growth model. This sector requires real factor inputs which are described by a financial production function. The model attempts to explain the determination of the capital intensity, the marginal product of capital, and per capita output when a financial sector gathers savings from surplus units and places these savings with deficit unit firms.

WILLIAM T. STANBURY, Ph.D. California (Berkeley) 1972. Changes in the size and structure of government expenditure in Canada, 1867-1968.

STEVE B. STEIB, Ph.D. Iowa State 1972. The Euro-dollar market as a source of U.S. bank liquidity.

The literature dealing with Euro-dollar borrowings has lacked theoretical foundations. In this dissertation a modern portfolio theory is used to derive the determinants of Euro-dollar borrowings by U.S. banks. The theory is empirically tested using the period from January 1965 to April 1971. It is indicated that Federal Reserve regulations are significantly associated with the levels of Euro-dollar borrowings.

ROBERT E. STERNENBERG, Ph.D. Houston 1972. A study of full costs versus marginal cost pricing of proposals to government agencies and their impact on contract awards and performance.

LI-TEH SUN, Ph.D. Oklahoma State 1972. Incidence of Montana state and local taxes.

The present study attempts to determine "who in the final analysis pays how much of the state and local taxes in Montana." The findings for the most part run counter to those of other similar studies. In these studies, the overall incidence of state and local taxes was found to be regressive. The overall incidence of Montana state

and local taxes is progressive before reaching the highest adjusted gross income bracket.

RONALD J. SUTHERLAND, Ph.D. Oregon 1971. Commercial bank portfolio behavior: A micro-economic analysis.

FREDERICK E. TANK, Ph.D. Wayne State 1972. Portfolio selection: A simulation study of the tradeoff between risk and consumption.

JOHN A. TATOM, Ph.D. Texas A&M 1971. Transactions costs and the supply of real average demand deposits. The role of banks in determining output is emphasized using a micro-economic model, where a firm produces real average demand deposits providing "checking" services in a competitive market. The bank, maximizing profits, faces a repurchase clause constraint that requires in addition to solvency, and reserves of currency, that it supply all transactions demanded by depositors. The model determines a service charge on transactions and rate of interest paid on average deposit balances. Reduced form equations are estimated for the period 1950-68. The results are somewhat surprising but are consistent with the theoretical structure.

GERALD R. THOMPSON, Ph.D. Virginia 1972. Expectations and the greenback rate, 1862-78.

RONALD G. TROSTLE, Ph.D. Kansas State 1971. An analysis of alternative tax sources to finance local services in Kansas.

I. LANG TSOUR, Ph.D. Oklahoma State 1972. Estimates of a dynamic demand for money model with distributed lags and autoregressive errors.

This study extends and refines estimation of the demand for money. In existing models, the demand for money is usually estimated by least squares. This method may result in inefficient estimates. The purpose of this is to remedy these deficiencies by applying the modified Gauss-Newton method of non-linear regression to a distributed lag model which contains two lag parameters and autoregressive error of the demand for money.

ROBERT A. VAN ORDER, Ph.D. Johns Hopkins 1972. Theoretical models of macroeconomic policy.

GERARD VILA, Ph.D. Harvard 1972. Does money matter? How much does it matter? How does it work? A selective and critical survey of recent developments.

This investigation of the "strength-of-money" debate concludes: The causal influence runs from "monetary policy" to income and not inversely. The results derive from pairwise causality tests between income and various monetary variables. They suggest that monetary policy works through disruptions in the flow of funds, not through the "quantity of money." Previously reported, insignificant fiscal multipliers (for example, St.

Louis Fed.) reflect statistical problems rather than the weakness of fiscal policy. Fiscal spending and monetary policies are the major causes of the cycle.

PAUL A. WACHTEL, Ph.D. Rochester 1972. A study of household investment.

BARRY WELLER, Ph.D. Pennsylvania State 1971. Fluctuations in municipal capital expenditures in the postwar period.

SHIRO YABUSHITA, Ph.D. Yale 1972. Essays on money and economic growth.

### International Economics

GEOFFREY ANDRON, Ph.D. Chicago 1972. A Heckscher-Ohlin trade model with transport costs.

KAZUKO K. ARTUS, Ph.D. California (Berkeley) 1971. An empirical examination of the Mundell model.

LEE D. BADGETT, Ph.D. Yale 1971. The response of processing activity to preferential tariff reductions: The Philippine case, 1900 to 1940.

HARVEY E. BALE, Ph.D. Maryland 1972. The role of price and redistributional factors in the adjustment to exchange-rate devaluation.

Recent theoretical and empirical discussions of the devaluation model have emphasized the importance of the impact of devaluation-induced short-run changes in the distribution of income on the level of consumption and import expenditures. A model of the redistribution effect is tested in this paper for three devaluation cases. The findings do not establish an important role for this effect, but rather confirm the importance of stabilization policies and the relative price effects of devaluation.

RICHARD N. BARRETT, Ph.D. Wisconsin (Madison). The Brazilian foreign exchange auction system: Regional and sectoral protective effects.

RUSSELL S. BOYER, Ph.D. Chicago 1971. The dynamics of an open, monetary economy: Growth and the balance of payments.

The thesis formulates a growth model using the Uzawa-Penrose framework, "opened up" in the context of a small economy with mobile financial capital and immobile physical capital with the rest of the world. The author investigates the problems of growth and the stages of the balance of payments following up work by Frenkel, Halevi, Laffer, and Mundell. Some previous conclusions are found erroneous because the authors fail to distinguish between cross-sectional and time-series results.

- FIKRET CEYHUN, Ph.D. Wayne State 1972. Export-performance of *U.S.* manufacturing industries: An econometric study.
- CHOENG-HOY CHUNG, Ph.D. Wisconsin (Madison) 1971. Interregional and international economic analyses of the world feed grain economy in 1980 with emphasis on the *U.S.* North Central Region.
- JAE W. CHUNG, Ph.D. New York 1972. The impact of protection on the factor market.  
The purpose of the present study is to explore an important condition for the complete sustenance of the Stolper-Samuelson theorem by integrating the production and demand sides. An empirical study on the demand side of the scarce-factor-intensive imports of the United States from Japan is carried out by employing a separably additive utility function of the Stone-Geary form.
- MICHAEL P. CLAUDON, Ph.D. Johns Hopkins 1972. International trade and technology: Models of dynamic comparative advantage.
- RUDIGER DORNBUSCH, Ph.D. Chicago 1971. Aspects of a monetary theory of currency devaluation.
- FREDERICK J. EGGERS, Ph.D. North Carolina 1971. Interdistrict settlement fund data and regional balance-of-payment adjustment.
- ASIM ERDILEK, Ph.D. Harvard 1972. A general equilibrium model of international trade with an application to United States-Japanese trade.  
In this study a multi-country, multi-sector, and multi-factor general equilibrium trade model is constructed and empirically implemented to analyze the 1960 bilateral trade of the United States and Japan. The positive and normative aspects of the model are examined within the framework of international trade theory. The simulation results, obtained under alternative commercial policies, are interpreted in the light of the actual trade of the United States and Japan in 1957-63.
- MARK W. FRANKENA, Ph.D. Massachusetts Institute of Technology. Export of engineering goods from India.
- MICHELE FRATIANNI, Ph.D. Ohio State 1971. Bank credit formation, money supply processes, and monetary and fiscal policies in an open economy: The Italian experience, 1958-69.
- ALAN GUMMERSON, Ph.D. Wisconsin (Madison) 1972. The factor content of European economic community international trade: An empirical evaluation of several hypotheses.
- JAMES F. HALSTEAD, Ph.D. Stanford 1971. Balance-of-payments adjustment and economic development: Areas of recent settlement.
- KURT F. HAUSAFUS, Ph.D. Northwestern 1972. A portfolio approach to international short-term capital movements.  
A decision model is developed from which expressions are derived for the supply of *U.S.* liabilities to the United Kingdom and the demand for *U.S.* claims on the United Kingdom. Three major findings emerge. Rates of return and wealth enter as arguments in the supply function. The demand function has as arguments variables related to trade finance, and is independent of financial variables. The United States has a strong comparative advantage relative to the United Kingdom in attracting *U.K.* capital.
- FREDERICK S. HIPPLE, Ph.D. Southern Methodist 1972. The demand for international reserves.  
In the extensive literature on the demand for reserve holdings, the major studies conclude that governments determine their reserve stocks by a cost-benefit approach. The present study develops a general model which explains reserves demand in terms of eight exogenous variables such as wealth, external disequilibria, holding cost, and others. The estimation was done for sixty-one nations, covering the period 1960-65.
- TOMOTAKA ISHIMINE, Ph.D. Wisconsin (Madison) 1972. Tariff preference and the Okinawan sugar industry.
- BERNARD KEY, Ph.D. California (Berkeley) 1971. The role of foreign contributions in Japanese capital formation, 1868-1936, with special reference to the period 1904-14.
- ALAN P. KIRMAN, Ph.D. Princeton 1971. Optimum tariffs in a general neoclassical model of international trade.  
The thesis develops a general version of the neoclassical model of international trade. Within this framework the problem of the optimum tariff is studied, both with and without retaliation. Particular emphasis is placed upon the nature of the assumptions, made implicitly in the traditional geometric analysis, which are required to obtain results in the general model.
- ROBERT C. KOHRN, Ph.D. Notre Dame 1972. International trade in rubber and rubber products: The *U.S.* demand for imports.  
The quantities of imports of tires and tubes, rubber and canvas footwear, synthetic rubber, and natural rubber were analyzed for 1930 through 1969 and for each of the decades in terms of price variable, a credit or income variable, an inventory variable, and a variable reflecting commercial policy. The effect of the consumer credit variable which was included as a surrogate for personal income appeared especially important in much of the analysis.
- JOHN KYLE, Ph.D. Wisconsin (Madison) 1972. The balance of payments in a monetary world.
- ARTHUR B. LAFFER, Ph.D. Stanford 1971. Private short-term capital flows.

WILLIAM N. LAYHER, Ph.D. Wisconsin (Madison) 1972. An analysis of foreign exchange rates and international short-term capital.

MARK K. LOKEN, Ph.D. Duke 1972. The impact of effective commercial policy on patterns of Canadian exports.

JACK A. LUCKEN, Ph.D. Boston College 1972. Interdependence and economic policy: The case of Japan and the United States.

This study describes the construction of a small linked model of the United States and Japan, and the results of policy experiments performed on it. The experiments reveal the short-run demand-induced effects of shifts in *U.S.* economic policy on the Japanese economy and their combined direct and indirect domestic effects. The variables chosen as policy instruments include receipts of personal, indirect, and customs taxes, changes in government consumption and in contributions for social insurance, and the discount rate of the Federal Reserve System.

DERMOT F. McALEESE, Ph.D. Johns Hopkins 1972. Import demand, protection, and the effects of trade liberalization in the Irish economy.

FRANCIS McCORMICK, Ph.D. California (Berkeley) 1971. The theory of forward exchange and capital flows under crawling exchange rates.

CARL H. McMILLAN, JR., Ph.D. Johns Hopkins 1972. Aspects of Soviet participation in international trade.

PAUL M. R. MALIMBA, Ph.D. California (Los Angeles) 1972. An economic evaluation of the international coffee agreement of 1962 and 1968.

The dissertation examines the contribution of post-war international commodity agreements to economic development of developing countries, using coffee as a case study. It found that the coffee agreements of 1962 and 1968 have, in the short run, benefited both exporting and importing members through a greater degree of market stability, but their contribution, in the long run, was uncertain in view of the economic shortcomings of their provisions.

DAVID J. MORAWETZ, Ph.D. Massachusetts Institute of Technology. Economic integration among less developed countries with special reference to the Andean group.

RICHARD S. MORELAND, Ph.D. Duke 1972. The impact of effective commercial policy on patterns of Canadian exports.

ALBERTO R. MUSALEM, Ph.D. Chicago 1971. Demand for money and balance of payments: The experience of Colombia, 1950-67.

The contribution is the identification as an alternative asset to real cash balances of those assets whose monetary yield is the expected rate of change in the price of goods related to international trade. Results

support the hypothesized speculative shifts between money and stocks of traded goods. The stock component of the balance-of-payments deficit is quite significant in some years, bringing some light into policies directed to solve the balance-of-payments crises.

NARAYAN K. NARGUND, Ph.D. Rutgers 1971. An econometric study of Indo-*U.S.* trade.

Using annual data (1951-68), Indo-*U.S.* import and export demand equations for major groups and sub-groups of commodities have been estimated. The study revealed that India's imports from the United States are, in general, more income elastic than her exports to that country and there exists an effective price competition in both the *U.S.* and Indian markets. Moreover, this study has incorporated the role of *U.S.* grants and aid in measuring India's demand for *U.S.* products and found it to be significant. Policy implications have also been analyzed.

CHIN K. PARK, Ph.D. California (Los Angeles). The role of the exchange rate in a developing economy: A case study of Korea.

The overvaluation of the domestic currency in Korea, forced mainly by short-term pressures has led to a successive deterioration in the "structure" of external transactions, and in the internal-external relationship. This study evaluates the potential role of the exchange rate as a policy variable required for "fundamental" adjustment of the balance-of-payments difficulties. It is shown that a 1 percent devaluation of the domestic currency is not only capable of improving the currency trade-balance by nearly 4 percent, but also compatible with the long-run policy goals of the economy.

MICHAEL G. PORTER, Ph.D. Stanford 1972. International interest rate differentials as behavior towards change rate expectations.

Differences in term structures of interest rates between countries with integrated capital markets are seen as reflecting expectations regarding the time path of the exchange rate. The hypothesis is tested for Canada under flexible and fixed rates and provides interesting results. A model of interest rate linkages under exchange rate uncertainty is incorporated. Exchange rate expectations are viewed as ultimately stemming both from uncertain future production and consumption and divergent national rates of monetary expansion.

ALOK RAY, Ph.D. Rochester 1972. Trade, protection, and economic policy: Essays in international economics.

PAUL N. ROY, Ph.D. Johns Hopkins 1972. The effects of changes in government monetary and fiscal policies in an open-economy one-sector monetary growth model.

PAUL SCHNITZEL, Ph.D. New York 1971. The Euro-dollar market: Some conceptual problems and a neglected function.

The author wishes first to clear up such problems as: how does a dollar come to be a Euro-dollar and given an agreeable definition of what a Euro-dollar is, how does one measure their growth? By way of addition, the author explores in a tentative way the relationship between the Euro-dollar market and the financing of U.S. exports. The method used is empirical and historical.

WAYNE M. SIMON, Ph.D. Iowa 1972. Foreign divestment under ANCOM.

This study concerns effort of the countries comprising the Andean Common Market to achieve rearrangement of investment and ownership patterns through divestment. The prospect exists of emergence of an important new investment form in the region—the consortium arrangement. Evidence points to willingness of many foreign firms to initiate investment within the region despite divestment action. Reasoning is presented to support the contention that divestment is not incompatible with economic integration and economic development.

LEAH T. J. SMITH, Ph.D. Johns Hopkins 1972. An empirical evaluation of import substitution in five large Latin American countries.

FELIPE M. SUVA, Ph.D. Massachusetts Institute of Technology. U.S. direct investments in the Philippines.

ROGER M. SWAGLER, Ph.D. Ohio State 1971. An analysis of variations in alternative terms-of-trade measures: A case study of Bolivia, 1925–65.

SUBIDEY TOGAN, Ph.D. Johns Hopkins 1972. Trade, growth, and international investment: A study in comparative dynamics.

MARY F. VAN LOO, Ph.D. California (Berkeley) 1971. The effect of direct investment on Canadian investment.

K. THOMAS VARGHESE, Ph.D. Wayne State 1971. Euro-dollar system, international stability of the dollar and measures of balance in U.S. international transactions.

CONSTANTIN S. VOIVDOAS, Ph.D. Columbia 1971. Exports, foreign capital inflow, and economic development.

The relationships investigated are those between exports, foreign capital inflow, and the rate of growth of total product in the light of the two-gap model. Empirical evidence supports the hypothesis of a positive relationship between exports and the rate of growth. It does not support the hypothesis of a positive relationship between foreign capital inflow and growth. The reason for the last result is a positive relationship between the former variable and the incremental capital-output ratio.

OBIE G. WHICHARD, Ph.D. North Carolina 1972. A cross-sectional investigation of the determinants of recent export performance of thirty-nine less developed countries.

This study attempts to analyze the determinants of export performance of less developed countries. In designing the analysis, a cross-sectional approach was selected on the basis of statistical considerations and a desire for generality. Within this framework, export performance is analyzed focusing on four explanatory factors: the structure of exports by-commodity group and by destination; the economic structure of the exporting countries; the existence of preferential trading arrangements; and per capita GNP.

JANET L. YELLEN, Ph.D. Yale 1971. Employment, output, and capital accumulation in an open economy: A disequilibrium approach.

### Business Administration; including Business Finance and Investment, Insurance, Marketing, and Accounting

ROLPH E. ANDERSON, Ph.D. Florida 1971. Consumer dissatisfaction: The effect of disconfirmed expectancy on perceived product performance.

In determining the effect on perceived product performance of disparity between consumer expectations and actual product performance, four psychological theories were considered: cognitive dissonance (assimilation); contrast; generalized negativity; and assimilation-contrast. Each theory makes a different prediction regarding disconfirmed expectancies. Major conclusions are that a substantial gap between high expectations and actual product performance may cause an unfavorable evaluation of a product; and that product evaluations are higher when objective product information, as opposed to little or no information, is provided customers.

TIMOTHY BATES, Ph.D. Wisconsin (Madison) 1972. An econometric study of Black capitalism: Feasibility, profitability, and financial soundness.

Criteria are presented for estimating the profitability of Black firms having various levels of sales, labor input, and capital input. Results of linear regression models suggest that profitability can be accurately predicted and that many Black-owned firms are economically viable. Discriminant analysis results indicate that credit evaluation criteria which are appropriate for judging white borrowers do not effectively identify credit risks among the population of Black borrowers.

JACK L. BISHOP, JR., Ph.D. Illinois (Urbana) 1972. A comparative evaluation of optimal inventory control policies.

Historical transaction history of thirty products stocked at seven locations was used to test the hypo-

thesis of Poisson generated sales. The assumed Poisson structure was then used to generate demand for a model of the production and distribution sectors of the firm. This model was used in the evaluation of the effect of alternative planning horizons, review periods, and inventory control doctrines on the measures of cost and performance.

DANIEL B. BOSSE, Ph.D. St. Louis 1971. A study of strategies of successful merchants in various retailing fields.

The role of strategy in growth and success was analyzed in divergent lines of trade, along with methods of origination. Conclusions indicated that retailing success was not contingent upon either distinctive external or internal strategies. Also, varieties of strategies were employed with equal effectiveness. Parental guidance and training was found to be more important as a success factor than financial assistance among the retailers studied. Strategy development was found to be an evolutionary and centralized process. Primary contributions include a model of strategy formulation for retailers.

GUY J. DE GENARO, Ph.D. Florida 1971. A planning-programming-budgeting system (*PPBS*) in academic libraries: Development of objectives and effectiveness measures.

The study explores the potential of *PPBS* as a vehicle for facilitating resource allocation decisions in academic libraries and examines some of the key problems involved in designing such a system. It formulates objectives and examines the problem of defining and measuring library benefits. Proxies for library benefits are presented and analyzed, along with their associated objective statements and measures of effectiveness.

CHARLES N. DENNIS, Ph.D. Arkansas 1972. An investigation into the effects of independent investor relations firms on common stock prices.

This investigation employs Myron Gordon's regression model, known as *ADLEV* model. Regression equations for twenty-three pairs of before-and-after samples (before and after the corporation hired an independent investor relations firm) were tested for coincidence using the method developed by Gregory Chow. No consistent effect by investor relations firms on the client corporations' common stock price was found.

CHARLES E. FISK, JR., Ph.D. Rice 1972. Simulated financial controls of research.

This study develops an algorithm for a class of non-linear assignment problems, and incorporates the algorithm into a digital simulation model for evaluating adaptive controls over research. Results from the model corroborate several unconventional hypotheses, notably that a simple rule of thumb for allocating resources among research projects works nearly as well as a theoretically optimal procedure, and that research firms—unlike insurers—decrease the predictability of returns by underwriting more risks in the same class.

GENE W. GRUVER, Ph.D. Iowa State 1972. Economic decision models under linear methods of decentralization.

First, a model of linked subunits is studied in the framework of a decomposable linear program. Extensions of the pricing and allocation rules are formulated to make the solution process more efficient. Second, computation of efficient output vectors for a university is discussed and a method for computing all efficient extreme points adjacent to a given efficient extreme point is presented. Finally, the relation between efficient production and noncompetitive price setting is studied.

ABRAHAM P. J. IMMELMAN, Ph.D. Florida 1971. An inquiry into the concept of public interest and its relevance to accounting theory and practice.

An inquiry into the meaning of the "public interest" concept and into the functions of accounting in a free enterprise market economy leads to the conclusion that the public interest concept is relevant to accounting theory and practice. Explicit recognition of the public interest as an implicit environmental accounting postulate accentuates the need to define it and explicit recognition of the public interest standard in public accounting practice accentuates, *inter alia*, the auditor's core function in society.

RAJINDER S. JOHAR, Ph.D. Wayne State 1972. Corporate investment and financial behavior: India, a case study.

A. WAYNE LACY, Ph.D. Iowa State 1971. The determinants of trade credit.

The hypothesis of this study is that the level of trade credit is influenced by sales risk considerations, which promote the granting of trade credit, and financial risk considerations, which discourage it. Empirical support is given that the trade credit decision is passive during periods of monetary ease but is affected by liquidity considerations during monetary constraint. Strong evidence is presented that the restraint discriminates against smaller firms, despite credit redistribution from larger to smaller firms.

WILLIAM R. POLLERT, Ph.D. Florida 1971. Managing differentiation and integration in fully integrated steel firms.

This study focuses on the effect of corporate-divisional differentiation, division performance, and managerial effort devoted to integration on the actual level of corporate-divisional integration achieved in steel firms. The findings demonstrate the need for steel management to be aware of the effects of their decisions on the determinants of effective conflict resolution and joint decision making.

RICHARD R. SPIES, Ph.D. Princetone 1971. Corporate investment, dividends, and finance: A simultaneous approach.

The purpose of this work is to investigate the capital budgeting decision of corporations, i.e., the determination of dividends, long- and short-term investment, debt and equity financing. Under the assumption that

the corporation's objective is to maximize the price of its common stock, a present-value model is used to derive the optimal capital budget. A complete partial adjustment system is then proposed and estimated for all manufacturing corporations and for ten industry subsectors.

ROBERT D. ST. LOUIS, JR., Ph.D. Purdue 1972. Mathematical programming and decentralization.

Mathematical programming is used as a tool to determine whether or not suboptimization is a necessary consequence of divisional autonomy and/or fully allocating overhead costs within decentralized firms. The Marshallian and Walrasian tatonnement processes as simulated via the Abadie-Sakarovitch and Dantzig-Wolfe decomposition algorithms are compared and contrasted. Overhead costs are integrated into the analysis by means of an extended Kaplan-Thompson type procedures. The demonstrated results are then generalized to a socialist economy.

PETER TREADWAY, Ph.D. North Carolina 1971. An explanation of the conglomerate merger movement—the conglomerate as a result of forces arising in the capital markets.

This dissertation examined the hypothesis that acquisitions of conglomerates tended to have a lower expected return (allowing for risk) before merger under their premerger managements. Premerger stock market and book returns for a sample of eighty-six acquired firms had statistically lower premerger returns. In addition, comparisons between the conglomerate and acquisition samples showed the conglomerates had significantly higher leverage and price/earnings ratios.

BERNARD YON, Ph.D. Cornell 1972. Advertising expenditure, time optimization, and related firm behavior.

A model is developed to identify the optimum level of advertising expenditure and length of time before a new advertising message is required. Estimation results give grounds for concluding that the lagged sales response to advertising is not geometric and the division of total advertising expenditure into two parts is a useful analytical service. As the lag structure could not be identified, optimization was determined with a replacement model based on a Markovian determination process.

### **Industrial Organization and Public Policy; including Economics of Technological Change, and Industry Studies**

ROBERT E. BABE, Ph.D. Michigan State 1972. The economics of the Canadian cable television industry.

ALAN R. BECKENSTEIN, Ph.D. Michigan 1972. An optimization approach for evaluating multiplant scale economies.

A theoretical framework is developed for investigating the extent of multiplant scale economies from optimal product specialization. A plant location model is used. Some new non-convex minimization techniques are applied to solve the non-linear programming problems. Empirical tests of multiplant scale economies in several industries are made. Also observed are the pure scale economies from large market share. The quantitative results from this analysis are essential in framing sound deconcentration policies.

HARRY BLOCH, Ph.D. Chicago 1971. Advertising, competition, and market performance.

This paper examines the implications for the compatibility of advertising and competition, of the relationship between advertising and market performance. For a group of food manufacturing firms, it is shown that the relationship between profit rates and advertising intensity is consistent with competition. On an industry study level, it is shown that the relationship between advertising intensity and three measures of performance, profit rates, economies of scale, and price-to-marginal-cost-ratio is consistent with competition.

BELINA B. CARR, Ph.D. Massachusetts Institute of Technology. A study of large mergers, 1965-69.

JOHN C. DANIELSEN, Ph.D. California (Berkeley) 1972. The determinants of inventive activity.

GERALD K. DAVIES, Ph.D. Washington State 1972. Rates and costs of grain transportation by railroad.

The purpose of this study is to reexamine the economic rationale behind the use of discriminating pricing in the railroad industry and to indicate the multitude of factors that must receive consideration in analyzing the extent of discrimination in a given rate structure.

HENRY G. DEMMERT, Ph.D. Stanford 1972. An economic analysis of the professional team sports industry in the United States with special emphasis on the Player Reserve System.

The professional sports industry, due to historically privileged status under the antitrust laws, is characterized by a unique set of institution arrangements endowing it with considerable monopoly/monopsony power. This thesis evaluates the economic impact of these institutions as well as their conventional justifications. The study concludes that the present institutional structure of the industry is economically unjustifiable. Economic performance, in terms of efficiency, equity, and product quality, could be improved by increased economic competition in both input and output markets.

BANARSI D. DHAWAN, Ph.D. Washington 1972. Economics of satellite television for India.

ERWIN DREESSEN, Ph.D. California (Berkeley) 1972. Elasticity of demand for labor: A cross-section study of wood products industries.

KEITH D. EVANS, Ph.D. Washington 1971. A comparison of the market structure, conduct, and performance of the drycleaning industries in Los Angeles and Seattle, with emphasis on the years, 1948-67.

LAWRENCE GOLDBERG, Ph.D. Brown 1972. The demand for industrial *R&D*.

Assuming that the output of *R&D* is uncertain a priori and that *R&D* is an investment shifting future product demand or supply, it is shown analytically that *R&D* per project depends upon the firm's cost of generating new information, its cost of future output, and its marketing ability. This theory and supporting econometric analysis of data from the drug industry explain the observation that *R&D* per unit of sales is lower for larger firms.

LAWRENCE G. GOLDBERG, Ph.D. Chicago 1972. The effect of conglomerate mergers on competition.

The major impetus behind recent proposals to restrict conglomerate mergers has been the allegation that conglomerate mergers have an adverse effect on competition. Little evidence has been collected to support this view. This paper examines the effects of conglomerate mergers upon market shares of acquired companies and upon concentration ratios within industries in which firms have been acquired in conglomerate mergers and concludes that conglomerate mergers are not a serious threat to competition.

JAMES R. GREEN, Ph.D. Oklahoma State 1972. The welfare effects of an antisubstitution law in pharmacy on the state of Oklahoma.

Because of an antisubstitution law, pharmacists can't substitute one of the lower priced versions of a drug on a prescription calling for a higher priced version of the same drug. As a result there has been a loss in the economic well-being of the consumers of prescription drugs. The welfare loss to consumers of Oklahoma for a one-year period is estimated to be one and one-half million dollars on eleven drugs.

JULIAN M. GREENE, Ph.D. Minnesota 1971. An empirical analysis of rate-of-return regulation and its effect on investment in steam-electric generation.

This study rejects the hypothesis that rate-of-return (*ROR*) regulation causes capital intensive electricity generation. VES plant production function parameters are estimated from a static cost-minimization *ROR* constrained model, best fit regressions occurring for a zero value of the *ROR* associated Lagrange multiplier. Capital-fuel substitution elasticity (0.3) and returns-to-scale (1.1) estimates are obtained. An appendix shows how annual load curves can facilitate estimation of instantaneous functions.

RAMON HAREL, Ph.D. Harvard 1972. Road user charges policy: The case of Israel.

The thesis deals with theoretical and empirical analysis of road user charges. The analytical aspects of the problem are devoted mainly to financing and payment schemes for road transport fixed facilities and determination of optimal charges which consider congestion

costs. In the empirical analysis, marginal social and private costs of road transport in Israel are estimated. A practical and consistent tax scheme is proposed which enables payment for road investments and leads to convergence of private and social costs. This scheme is integrated with the existing indirect tax structure of the country.

DENNIS L. HEFNER, Ph.D. Washington State 1971.

The relationship between company profits and market structure in selected food processing industries.

Regression analysis within a discrete industry is employed in an examination of the relationship between company profits and aspects of market structure and experience in five food processing industries. The study is confined to the eight largest registered companies of five food processing industries in each year during the period 1949-67.

ANTHONY F. HERBST, Ph.D. Purdue 1971. Trade credit in the U.S. lumber and wood products industry.

An investigation into the determinants of trade credit at the industry level of aggregation, where a void had existed. The specific question considered was that of the extent to which observed changes in the trade credit variables are endogenous phenomena within the industry system of economic relationships. Factor analysis was employed to summarize separately the variance in industry and in macro-economic variables in support of the results obtained with alternative regression models.

GEORGE E. HOFFER, Ph.D. Virginia 1972. Physician-ownership in pharmacies and drug repackagers.

FRANK E. HOPKINS, Ph.D. Maryland 1971. Transportation cost and industrial location: An analysis of the household furniture industry.

Transportation cost has been a major concern of industrial location theory. Unfortunately, recent empirical studies have either neglected or have used proxy variables to account for the influence of transportation cost upon industrial location. The model developed in the dissertation uses shadow prices from linear programming transportation problems and other traditional variables to explain the location of the household furniture industry at the county level in the United States between 1964 and 1966.

MICHAEL S. HUNT, Ph.D. Harvard 1972. Competition in the major home appliance industry, 1960-70.

Why did the major home appliance industry, a highly concentrated industry with high barriers to entry, experience low profits, declining prices, and continual product innovation throughout the 1960's? This thesis, based on extensive field research, argues that economic and managerial differences between the major competitors led them to view different prices and *R&D* efforts as optimal and prevented them from reaching a tacit agreement that would have allowed them to exploit the potential monopoly profits.

- BRUCE L. JAFFEE, Ph.D. Johns Hopkins 1972. Aspects of the regulated public utility: Misallocation, marginal cost pricing, and depreciation.
- JAMES JONDROW, Ph.D. Wisconsin (Madison) 1972. A measure of the monetary benefits and costs to consumers of the regulation of prescription drug effectiveness.
- PAUL L. JOSKOW, Ph.D. Yale 1972. A behavioral theory of public utility regulation.
- KIYOSHI KAWAHITO, Ph.D. Maryland 1971. The steel import problem of the United States and the Japanese steel industry.  
The growth of U.S. steel imports from Japan between 1958 and 1968 is analyzed with an econometric model based on the differential equation of the Gompertz curve. Import projections up to 1975 under various assumptions are added. The thesis also presents a survey of the recent history of the Japanese steel industry and analyzes its various facets today as compared with the American and European counterparts.
- THEODORE E. KEELER, Ph.D. Massachusetts Institute of Technology. Resource allocation in intercity passenger transportation.
- WILLIAM C. KERBY, Ph.D. Oregon 1971. Influence of industrial structure on the profit rates of large manufacturing corporations, 1954-65.
- DAVID E. KUNKEL, Ph.D. Wisconsin (Madison) 1972. Market structure, conduct, and performance: The Turkish cotton textile industry as a case study.
- RICHARD W. LICHTY, Ph.D. Kansas State 1971. An analysis of alternative fresh and frozen meat distribution systems.  
The research investigated whether there were significant cost differences between cutting meat at a central location as opposed to a retail store. Costs of preparing frozen brand name products were also investigated. An input-output production function, using primary data, was used to evaluate the various systems. It was found that significant cost savings were possible with central production but not necessarily through freezing.
- MARILU H. MCCARTHY, Ph.D. Georgia State 1972. A market and input shares approach to the spatial distribution of manufacturing.  
This paper develops a procedure for measuring regional shares of manufacturing industry based on material input supply and demand for output within the region. Regional shares of supplying or purchasing industries were weighted with technical coefficients defining their relative importance to users of intermediate or final output. The summation of all weighted output shares of supplying industries is one measure of the region's potential for production in a particular manufacturing industry. Similarly, the summation of all weighted shares of market demand from industrial manufacturers, consumers, and government is a measure of potential successful operation for particular manufacturing industries.
- JAMES R. MCCAUL, Ph.D. Maryland 1971. The direct maritime subsidy program: An analysis and a proposal for change.
- MICHAEL J. MAGURA, Ph.D. Boston College 1972. An empirical investigation into a non-linear advertising-concentration relationship.  
The purpose of this study is to determine whether advertising has a statistically significant impact on levels of industrial concentration. Theory suggests that, if a relationship exists, it will be of a non-linear form—such as the log-reciprocal, semi-log, or double-log—rather than the linear assumed in previous empirical studies. Data were compiled for 118 firms aggregated into 34 four-digit SIC industries for 1954, 1958, 1963, and 1966.
- JAMES B. MARSCH, Ph.D. Chicago 1972. A theory and estimation of gold supply: The case of South Africa, 1956-66.  
The study integrates the theories of multiple outputs, natural resource exhaustion, and gold supply into a general, theoretical model of mining production. The model is then tested against data from the South African gold-uranium industry, and short-run gold supply elasticities are computed.
- R. CHARLES MOYER, Ph.D. Pittsburgh 1971. A model of the determinants of registered bank holding company acquisitions.  
The dissertation examines the reasons behind recent expansion of registered bank holding company (RBHC) activity. The basis of the model developed is the hypothesis that growth-oriented RBHCs seek to maintain or increase their control of banking deposits, relative to other competing banking organization. This growth orientation is constrained by the legal and regulatory environment, resource limitations, and constraints related to size and location.
- RICHARD W. NELSON, Ph.D. Yale 1971. Regulating technical change: The case of communication satellites.
- NORMAND R. NOEL, Ph.D. Boston College 1972. Market structure and the degree of rationalization.
- RICHARD B. NORGAARD, Ph.D. Chicago 1971. Output, input, and productivity change in U.S. petroleum development: 1939-68.  
New technology and declining resource quality have had large impacts on petroleum development between 1939 and 1968. Without new technology, real well costs would have increased 70 percent due to shifts in areas drilled, 45 percent due to drilling to deeper depths, and 35 percent due to the decrease in the success rate, and 233 percent due to the combination of these factors.

Because of new technology and cheaper inputs, real well prices increased only 64 percent.

GERALD T. O'MARA, Ph.D. Stanford 1971. A decision-theoretic view of the microeconomics of techniques diffusion in a developing country.

CENGİZ OZOL, Ph.D. Vanderbilt 1972. Monopoly regulation and a description of the behavior of electric utility firms: Some theoretical and statistical analyses.

JOHN P. PALMER, Ph.D. Iowa State 1971. The profit performance effects of the separation of ownership from control in large U.S. industrial corporations.

During the past forty years, the control of many large corporations has left the hands of the corporations' owners. This study first discusses the extent of the separation of ownership from control and some of the reasons for its importance. Then tests are performed on the effects of this separation on reported profit rates, the response of reported profit rates to tax rate changes, and the variation of profit rates over time. The results of these tests indicate that, generally, manager-controlled firms have a different profit performance than owner-controlled firms.

STEPHEN A. RHOADES, Ph.D. Maryland 1971. The effect of diversification on industry profit performance.

This study examines the proposition that the presence of conglomerate firms in an industry is a structural variable that influences industry profit performance. The study is based on a sample of 241 four-digit industries (1963). The results support the proposition under examination and suggest that diversified firms are relatively profitable in their primary industry but relatively unprofitable in secondary industries. These results are used to assess several hypotheses of the effects of conglomeration.

EDMUND J. SEIFRIED, Ph.D. West Virginia 1971. The determinants of price and rate of return for publicly owned electric utilities.

HERVEY I. SELEY, Ph.D. Columbia 1971. An *ex post* benefit-cost analysis of the federal subsidy to the shallow-draft water carrier industry.

An attempt is made in this study to determine whether the federal subsidy to the shallow-draft water carrier industry in the form of channel and harbor improvements and maintenance has been justified. A benefit-cost model applicable to such improvements is developed. We conclude that it is impossible at the present time to determine whether the subsidy has been justified but that continued utilization of the present shallow-craft system is warranted.

ROBERT SHISHKO, Ph.D. Yale 1972. An empirical study of technical change through product improvement.

KENNETH B. STANLEY, Ph.D. Michigan State 1971. Market structure and investment behavior in the international telecommunications industry.

AHMAD M. TABBARA, Ph.D. West Virginia 1972. The current merger movement and its effect on the level of industrial concentration.

HOSSEIN TAHMASSEBI, Ph.D. Indiana 1972. The impact of collective bargaining in international oil dealings: A case study of OPEC.

This study analyzes the characteristics of the international oil markets both before and after the establishment of the Organization of the Petroleum Exporting Countries (OPEC), taking up the relationships between the oil companies and the producing countries in both periods, with more emphasis on the post-OPEC period. The study indicates that through collective bargaining, the producing countries have been able to increase extensively their leverage in the international oil markets and their share of profit on each barrel of oil exported.

GEORGE W. THOMAS, Ph.D. Purdue 1971. Stochastic product differentiation.

EDUARDO J. TRIGO, Ph.D. Wisconsin (Madison) 1972. Structural changes in the food retailing market in the Buenos Aires metropolitan region of Argentina during the 1960-70 decade.

LARRY R. TRUSSELL, Ph.D. Arkansas 1972. An economic analysis of the relationship between the farm equipment manufacturer and his franchised dealer.

This dissertation analyzes the franchisor-franchisee relationship in the farm equipment industry. The framework is the permissible limits of franchisor control over the franchisee's operations; the industry's competitive structure; and current economic factors affecting industry performance. Data were gathered by a questionnaire survey of farm equipment dealers and short-line manufacturers. The study concludes that competition in the farm equipment industry would be substantially reduced if the long-line dealers were not permitted to handle short-line implements.

K. SRINIVAS UPADHYAY, Ph.D. Kansas State 1971. An analysis of the structure of the farm mortgage credit market in Kansas.

NAI-CHI WANG, Ph.D. Maryland 1971. A multistage linear programming model of transportation.

A new "sub-commodity" concept has been developed into a forecasting model to accommodate cross-hauling. All actual shipments included in the linear programming solution are "sub-commodity one." Existence of other shipments suggests that goods involved in such shipments are not analytically homogeneous with those in sub-commodity one. All other shipments included in the second linear programming solution are "sub-commodity two" and proceed on iterations to analyze what remains. The model has been tested with coal shipments.

ROBERT F. WARE, Ph.D. Michigan State 1972. A test of the entrepreneurial versus managerial hypothesis in the theory of the firm.

BERNARD M. WOLF, Ph.D. Yale 1971. Internationalization of U.S. manufacturing firms: A type of diversification.

### Agriculture and Natural Resources

THOMAS G. ANDERSON, Ph.D. California (Berkeley) 1972. A parametric programming approach to regional demand estimation for the allocation of water and water-related resources.

A. PAUL BAROUTSIS, Ph.D. Purdue 1972. Economic incentives in water pollution abatement.

METIN BERK, Ph.D. Iowa State 1971. Charging structure of Iowa farmland ownership.

This study is undertaken in order to identify and measure the factors influencing tenure of farmland owners and to determine the changing characteristics of Iowa farmland owners. Discriminant analysis is performed on four land tenure groups. The farmland owners most likely to change their tenure status are identified and the factors influencing their tenure classification are evaluated. The second objective of the study identifies the changing characteristics of Iowa farmland owners since 1946. The findings are presented and tested for statistical significance.

ERNEST W. CARLSON, Ph.D. Boston College 1972. Theoretical and empirical explorations in the economics of marine resources.

JOHN E. CREMEANS, Ph.D. American 1972. Pollution abatement and economic growth: An application of the von Neumann model.

Mathematical and modeling techniques useful in application of the von Neumann model of an expanding economy to pollution problems are developed. These include construction of pollution-growth tradeoff curves and the calculation of effluent charges. A test of a 92 process, 88 product system is based on 1957 input-output data and recent  $SO_2$  emission and abatement data. The results are promising, but no claim is made that they apply to current problems.

MICHAEL P. CUDDY, Ph.D. North Carolina State 1972. Apple production strategies for North Carolina growers.

This study estimates the expected returns and variance of returns to various apple production alternatives in North Carolina, given probable trends in relative prices of various types of apples. It establishes the profitability relationships for alternative apple varieties, tree types (standard and size controlled), and market outlets (fresh and processing markets). The most profitable production strategies for North Carolina apple growers are analyzed for a range of apple prices and production costs.

CARLOS T. DE ARRIGUNAGA, Ph.D. Texas A&M 1971. Trade, development, and structural change: The future of Mexico's henequen industry.

This study describes the structure of the world hard fiber industry and markets; the factors that lead to the decline of hard fiber markets; the past and present structure of the henequen industry, and how this structure relates to markets for henequen fiber and products. The determinants of the demand for henequen fiber and products are examined and the results of a statistical analysis are presented.

PAUL DUANE, Ph.D. North Carolina State 1971. Analysis of wool price fluctuations: An economic study of price formation in a raw material market.

The purpose of this study was to determine how wool prices are formed and how their variance has responded to competition between wool and other fibres. A seven-equation model was derived for estimation. The wool supply equation was estimated using ordinary least squares. The other equations were estimated by two-stage least squares using a limited number of principal components of the predetermined variables. It was demonstrated that the variance of wool prices falls significantly as the price elasticity of demand increases.

WALLACE C. DUNHAM, Ph.D. Cornell 1971. The role of national trade associations in the food industry.

A survey of the origin, functions, structure, and internal characteristics of trade associations in the food industry. Special attention is given to financing, control, and changing purposes over time. Certain opportunities for merger, consolidation, and joint venture are explored.

SOPHIA I. EFSTRATOGLOU, Ph.D. North Carolina State 1972. The economic effect of inter-county transfer of flue-cured tobacco quota.

The economic factors affecting the supply and demand of marketing quota were identified and used to estimate simultaneous and single (reservation demand) equations for production belts in six southeastern states. Farm labor market supply and demand equations were also estimated. Regression coefficients were used to estimate transfer quantities and equilibrium rental rates if transfer could be made on an intercounty (intrastate) and interstate basis.

MILTON H. ERICKSEN, Ph.D. Kansas State 1971. An analysis of how kinds and levels of specification and aggregation detail affect accuracy and usefulness of recursive programming estimates of production response.

The purpose of this dissertation was to evaluate methods of specification and aggregation relevant for development of recursive programming models used for predicting short-run aggregate production response of crops. The effect of different levels of specification and aggregation detail on the accuracy of production response estimates was analyzed. Four empirical programming models were developed to test alternative methods and levels of detail. The results showed only

slight differences in accuracy of production response estimates between models with and models without individual farm detail.

FERENC FEKETE, Ph.D. Iowa State 1971. Economics of cooperative farming: Objectives and optima in Hungary.

The study analyzes the historical development and socioeconomic determinants of cooperative farming and derives optimum solutions characterizing cooperative ends and capacities. Marginal analysis and programming models are concerned with the objectives and constraints of the large-scale collective enterprise, with those of individual cooperative households and with those of the cooperative organization as a whole. The study also presents some recent economic policy problems of Hungary in the light of the reform of the economic mechanism introduced in 1968.

TERRY A. FERRAR, Ph.D. Purdue 1971. Studies in environmental management.

JOSE M. FRANCO, Ph.D. Wisconsin (Madison) 1971. The legal insecurity of landed property in Venezuela: A case study of the registry and cadastral systems.

RICHARD F. FULLENBAUM, Ph.D. Maryland 1971. A general equilibrium demand model for living marine resources: An application of general equilibrium and common property resource theory to the U.S. seafood sector.

The purpose of this study is to extend the traditional model of common property resource exploitation to a more general equilibrium frame of reference. The most pronounced modification consists of the specification of cross partial derivatives of demand between the major species of seafood consumed in the United States. The general policy implication is that only slight, nonfundamental changes in the market mechanism are required to prevent excessive entry of capital and labor.

JOSE A. B. S. GIRAO, Ph.D. Cornell 1971. The impact of income instability on farmers' consumption and investment behavior: An econometric analysis.

Individual farm firm and household data for a period of seven years were classified into two groups based on degree of income instability. Econometric models of consumption and investment behavior were developed and estimated for the two groups. Variability in income apparently does not influence the average or long-run marginal propensities to consume but seems to influence investment behavior in a variety of ways.

THOMAS A. GRIGALUNAS, Ph.D. Maryland 1972. Waste generation, waste management, and natural resource use: An economic analysis of the primary copper industry.

GERALD A. HARRISON, Ph.D. Iowa State 1971. Alternative statistical models for estimation of acreage diversion with farm, county, and state levels of data: Feed Grain Program, 1961-70.

This study provides information to policy makers on factors influencing levels of acreage diversion from 1961-70 and models for predicting acreage diversion under the Feed Grain Program. Analysis is carried out on 1,449 observations from the 1962 Feed Grain Program Survey of the Corn Belt. Ordinary multiple regression, discriminant and limited dependent variable functions are fitted to the data for comparative analysis of model predictive ability.

PAUL L. HELSING, Ph.D. Washington State 1972. Fish vs. dams: The economics of maintaining the Columbia River Basin fishery.

The purpose of this study is to designate the requirements of economic efficiency and examine them in relation to the measurement of the economic and social benefits of the anadromous fisheries resource and the costs associated with providing fish passage in the Columbia River Basin.

L. DEAN HIEBERT, Ph.D. Wisconsin (Madison) 1972. Risk, tenancy, and resource allocation in low income agriculture.

HAROLD H. HISKEY, Ph.D. Utah State 1972. Optimal allocation of irrigation water: The Sevier Basin.

Four subbasin models were developed to generate value of marginal product schedules for water. There were indications of misallocations in this closed Utah River Basin. Reallocations involving 28 percent of the water would increase basin net farm incomes by approximately 10 percent. Since there are significant return flows, it was concluded that water allocation policies should be based on net stream depletion to prevent upstream water use efficiency improvements from depleting allocations to downstream users.

SIGMUND A. HORVITZ, Ph.D. Houston 1971. Economic optimal levels of control of sulfur dioxide emissions from the combustion of fossil fuels.

JAMES J. JACOBS, Ph.D. Iowa State 1972. Economics of water quality management: Exemplified by specified pollutants in agricultural runoff.

The thesis analyzes the role of economics in environmental quality management, with particular reference to the appropriate level of water quality in a selected use area. Sediment and phosphorus in agricultural runoff were the pollutants selected. By estimating the annual costs and returns and sediment and phosphorus losses, the minimum cost of achieving specified levels of water quality, via agricultural practices were obtained. In addition, possible means to determine benefits from water quality management were investigated.

GEORG KARG, Ph.D. Iowa State 1971. Optimal time paths of cattle and hog marketings.

Time paths under investigation were quarterly U.S. cattle and hog marketings. Efficiency criteria of time paths were farmers' annual cash receipts from cattle or/and hogs. To optimize the criteria for selected past and future years with respect to marketings, prices in-

cluded in cash receipts were replaced by price equations derived from a structural model of the beef and pork industry, and marketings were subject to various constraints. Past time paths were compared with optimal ones.

RONALD D. KAY, Ph.D. Iowa State 1971. A dynamic linear programming model of farm firm growth in North Central Iowa.

A dynamic linear programming model covering eight years was developed to study farm firm growth. Optimal growth strategy consisted primarily of increasing crop production by renting land and expanding swine production utilizing a capital intensive production method with the objective function being maximization of discounted net returns. The model relied heavily on this same strategy under different capital restrictions, interest rates, labor costs, and objective functions.

MARK C. KENDALL, Ph.D. Rochester 1972. Production externalities in the fishing industry.

CHARLES W. LANDRY, Ph.D. Arkansas 1972. Arkansas agricultural credit institutions and agricultural credit markets under conditions of continued economic growth.

The agricultural sector of Arkansas has experienced significant structural changes in recent years. There has been a decrease in the number of farm units and farm labor, along with increases in farm size. These changes have been accompanied by large increases in agricultural credit extended as well as increased reliance on institutional sources of funds. A test of the adequacy of Arkansas agricultural credit markets showed responsiveness to the increased credit demands placed upon them.

RUSSELL LIDMAN, Ph.D. Wisconsin (Madison) 1972. The distribution of benefits of U.S. farm programs: A case study for 1969.

HAROLD LOFGREEN, Ph.D. Iowa 1972. Economic analysis of alternative water pollution control measures.

This study explores the potential cost of commitments to water quality as well as ways in which that cost might be minimized within federal and state government guidelines of improved water quality. Cost estimates of controlling industrial and municipal pollution are developed from engineering sources for a number of treatment sequences which are used as proxies for varying levels of water quality. The estimates of costs of capital facilities, operating costs, and average costs are made for the United States and Iowa.

JEANNE M. MCFARLAND, Ph.D. California (Berkeley) 1972. California municipal solid waste management: A case study in public enterprise.

HOWARD C. MADSEN, Ph.D. Iowa State 1972. Future water and land supply-demand balances for U.S. agriculture: A spatial analysis incorporating producing areas, consuming regions, and water supply regions.

This study determines whether the nation has enough water and land to satisfy its future food and fiber needs. A large-scale linear programming model of U.S. agriculture incorporates alternative sets of parameters for population, water prices, technological advance, exports, and absence or existence of government supply control programs in the year 2000. A general conclusion is that projected domestic food and fiber and export demand will not press against available water and land resources in 2000.

RICHARD J. MIKES, Ph.D. Iowa State 1971. An appraisal of Iowa's grain elevator industry and potential structural adjustments.

The study analyzed the current elevator structure, potential savings in a modified structure with larger elevator, and potential structural adjustments over time. Estimated economies of scale were based on both a statistical analysis of costs in elevators and an engineering simulation. Crop production, livestock production, grain movements, harvesting technology, number of elevators and average size of elevators were projected for the nine crop reporting districts in Iowa.

MUHAMMAD NAZIR, Ph.D. Utah State 1972. Economic efficiency of grazing systems.

From the viewpoint of the Bureau of Land Management (BLM) as a proprietary agent, the internal rate of return on grazing system investments was 2.37 percent. From the point of view of cattle ranchers grazing BLM ranges, grazing systems increased aggregate profits by 3.65 cents per pound of beef sold. Federal investment in grazing systems was a sound venture, for the net revenue received by society in the form of added BLM grazing revenues and increased rancher profits totaled \$3.43 for each additional unit of forage produced. This is in addition to the increased nongrazing public benefits associated with increased forage production.

EMMANUEL O. OYINLOLA, Ph.D. Iowa State 1971. A confirmatory analysis of an exploratory factor analytic study of the fluid milk bottling firms in the north central region.

This is a follow-up of Oehrtman's dissertation. In order to test the validity of his exploratory factor analysis of the fluid milk bottlers, new confirmatory factor analysis techniques were developed. It was found that the results of this study were inconsistent with his solution. While the results of this dissertation were not generalized because of the limited sample size, the techniques developed shed new light on the use of factor analytic model in economic analysis.

DAVID E. PINGRY, Ph.D. Purdue 1971. Programming applications to the economic problems of water quality control.

EDMUNDO PRANTILLA, Ph.D. Iowa State 1972. Economic optimization models of multiple cropping system applied to the Philippines.

The study involves the construction of optimization models for multiple cropping operation. Two basic

linear deterministic models were formulated. The two models can ascertain in advance the optimal area allotted to each crop in every cropping period given the technological coefficients and the resource constraints. The optimal input requirements for the crops included in the optimal activity vector are on a monthly basis. The two models were applied using Philippine data on crop production.

RODOLFO QUITOS, Ph.D. Wisconsin (Madison) 1971. Agricultural development and economic integration in Central America.

DARYLL E. RAY, Ph.D. Iowa State 1971. An econometric simulation model of U.S. agriculture with commodity submodels.

An econometric simulation model was developed which causally links resource use, production, price, utilization, and income for major agricultural commodities. Based on this quantitative model, the implications of changes in selected variables on resource use, output, and income were investigated for individual commodities and U.S. agriculture as a whole. A modified version of the model was used to project resource requirements to 1980 under alternative 1980 production needs.

ROBERT O. RECK, Ph.D. Maryland 1971. The regional economic and environmental impacts of the uranium mining and milling industry.

This dissertation measures the external diseconomies of uranium mining and milling activities in the Colorado River Basin. Location quotients and regional input-output multipliers indicate the industry's economic importance to the region. Environmental effects are also summarized. Expected increases in radiological, cancer-related deaths are estimated and used in the computation of the costs of this disease. Losses due to premature burial, treatment, and absenteeism are also considered. In this region there is a need for a reallocation of resources consistent with the principles of life-cycle planning.

REFUGIO I. ROCHIN, Ph.D. Michigan State 1971. A micro-economic analysis of smallholder response to high-yielding varieties of wheat in West Pakistan.

This thesis documents a rapid diffusion and adoption of dwarf wheats among smallholder farmers of rainfed land in a densely populated district in Pakistan. "Innovation" and "awareness" are correlated against sets of communication and economic variables. The pattern of off-farm migration is studied. The impact of dwarf wheats is measured in comparison with traditional wheats by means of a Cobb-Douglas function and budgetary analysis. Recommendation for development strategy conclude the study.

JAMES G. RYAN, Ph.D. North Carolina State 1972. A generalized crop-fertilizer production function.

This study is concerned with the use and implications of soil testing for the purpose of making fertilizer recommendations for potato farmers in Peru. It establishes the quantitative relationship between soil characteris-

tics and fertilizer response and derives procedures for making recommendations to farmers considering their probable aversion to risky alternatives. The average value of soil test information is estimated, as is the probable impact of fertilizer adoption on prices and resource use.

EDUARDO SARMIENTO, Ph.D. Minnesota 1971. Efficient allocation of resources in the supply of water for domestic consumption.

STEPHEN J. SCHMUTTE, Ph.D. Purdue 1971. Interrelations of law and economics: The case of stream pollution.

This study explores Coase's analysis of social cost as a model of economic policy concerning stream pollution. Two theories of property rights used by courts in the past to adjudicate private pollution disputes are compared in terms of their impacts upon the efficiency of resource allocation. The study concludes with a discussion of arguments made against reliance upon a system of private, judicially enforced property rights and a tentative case for greater reliance upon such a system.

ARNDT SEIFERT, Ph.D. Michigan State 1972. The time price system: Its application to the measurement of primary outdoor recreation benefits.

ROBERT L. SPORE, Ph.D. Pennsylvania State 1972. Property value differentials as a measure of the economic costs of air pollution.

This study observes air pollution damage costs as they are reflected by differentials in the values and rents of residential properties in Pittsburgh, Pennsylvania. A partial equilibrium model of the residential property market is developed which recognizes the existence of submarkets within the overall property market of the metropolitan area. Besides measures of air quality, the estimated equation contains variables to account for variations in demand determinants among submarkets and neighborhood, locational, and physical attributes of residential properties.

THOMAS H. STEVENS, Ph.D. Cornell 1972. Equity and water resource development.

A conceptual model is formulated to analyze the distribution by income class of the net benefits arising from water and related land resource programs. The model is constructed to encompass cost-sharing procedures, different discount rates for the distributional classifications, and can be applied in a regional as well as a national context. It is empirically implemented for five Corps of Engineers multipurpose water resource projects, and the sensitivity of the empirical results to alternative incidence assumptions, discount rates, and data sources tested. Conclusions for planning and evaluation of alternative public investments are drawn from the empirical results.

FREDERICK J. WELLS, Ph.D. Massachusetts Institute of Technology. An economic evaluation of the U.S. desalting research and development program.

**Manpower, Labor, and Population;  
including Trade Unions and  
Collective Bargaining**

GARY L. APPEL, Ph.D. Michigan State 1972. Effects of a financial incentive on AFDC employment: Michigan's experience between July 1969 and July 1970.

LAWRENCE V. ASCH, Ph.D. North Carolina 1971. Selective variation of private labor demand: Rationale, practice and potential.

CORRY F. AZZI, Ph.D. Harvard 1972. Manpower programs and the public investment decision.

Manpower programs are intended to benefit the unemployed and the unskilled by altering employers' hiring, training, and promotion decisions. In general the programs do this by either directly or indirectly subsidizing employers. If the programs benefit mostly employers, the result could be a regressive income distribution from taxpayers to corporate shareholders. The thesis presents theoretical material which suggests that employers could be expected to be the primary beneficiaries of many manpower programs and then analyzes the distribution of programs' benefit within two firms.

DAVID E. BERGER, Ph.D. Washington (St. Louis) 1972. An analysis of growth in mature regions: A study in labor market dynamics.

Alternative explanations for economically mature regions were studied. The traditional export-base, industrial composition, and Borts-Stein neoclassical growth theories were examined. The first two had little explanatory power. The third yielded more satisfactory results; however, its shortcomings severely limited its usefulness. A neoclassical model on a disaggregate level produced a satisfactory theory of wage determination which was able to account for regional economic stagnation.

REMI BOELAERT, Ph.D. Wisconsin (Madison) 1972. Wage-price dynamics in EEC countries: An analysis of the unemployment-inflation tradeoffs in the Common Market.

LARRY K. BOND, Ph.D. Utah State 1972. An analysis of factors influencing family farm residence location.

Census data reveal that the percentage of farm operators in the United States that live off-farm is increasing. It has been suggested that this may be due to shifting of residence. What are the reasons farmers move into town? This study identifies and evaluates variables that influence residence location, determines to what extent residence shifting is actually occurring in Utah.

MICHAEL J. BOSKIN, Ph.D. California (Berkeley) 1971. The effects of taxes on the supply of labor: With special reference to income maintenance programs.

JOHN E. BROWN, Ph.D. North Carolina State 1972. A study of the economic variables affecting the valuation of a human life in legal decisions.

Legal statutes in this country have long provided that dependents of anyone wrongfully killed may sue the negligent party for the loss suffered as a result of the death. Traditionally, this loss has been primarily the "economic" loss. This research developed a model to explain and predict court awards in wrongful death cases. Economic and noneconomic variables were entered into the model. Dividing the data into a large and small income group demonstrated that the economic variables have much more explanatory ability in the instances of large incomes. The noneconomic variables are more important in the smaller awards.

WILLIAM R. BUECHNER, Ph.D. Harvard 1972. Technological change and the occupational composition of the American labor force, 1950-60.

This dissertation examines the effects of technological change on the occupational composition of the U.S. labor force between 1950 and 1960. An input-output model, with each industry's labor coefficient disaggregated into 240 occupational coefficients, is used to measure the effects of final demand changes and various measures of technological change on the economy's demand for labor by occupation. The model is also used to determine whether the occupational substitutions which occurred were affected by changes in the relative wages of the different occupations.

WINSTON C. BUSH, Ph.D. Washington (St. Louis) 1971.

A model of family size and intergenerational transfer and its implications for economic growth.

A model of family size and intergenerational transfers is constructed and its comparative static implications are investigated. Since both the determinants of family size and saving for intergenerational transfers are derived, the model is well suited for the study of the economic growth process. Therefore, results pertaining to economic growth are derived. These results are found to be consistent with past empirical finding in the area. Policy implications of the model are also investigated.

NANNETTE C. CITRON, Ph.D. Boston College 1972. The identification of an effective salary schedule for Massachusetts public school teachers.

WILLIAM D. COOK, Ph.D. Chicago 1971. The demand for contraceptive information, goods and services: An analysis of the Orleans Parish Family Planning Program.

The research consists of an analysis of data from an experiment in which women were required to pay \$10 for one year of contraceptive goods and services. A control group attended free of charge. Multiple linear regression analysis was used to estimate the probabilities of attendance, initiation of clinic methods, and participation (demand) and the impact of twenty variables on those probabilities.

ELIZABETH CROWELL, Ph.D. Indiana 1971. An analysis of discrimination against the Negro in the building trades unions.

An analysis of 1967 union membership data obtained from the Equal Employment Opportunity Commission

demonstrates a definite pattern of racial segregation within the building unions. Over 75 percent of all Black membership in these unions is concentrated in the lowest-paid, lowest-skilled union. In only four other unions is Black membership proportional to their representation in the labor force. It can be concluded that in ten unions—most of them the more desirable, higher paid trades—Blacks have been arbitrarily discriminated against.

ALVIN M. CRUZE, Ph.D. North Carolina State 1972. Determinants of curriculum offerings in the North Carolina community college system.

This research developed a model of curriculum offerings in the North Carolina community college system. Factors included were industry demand, student demand and internal operating incentives, with primary attention devoted to industry demand. The research extended Becker's theory of employer incentives to finance training by including the alternative of hiring a trained worker. Due to the dichotomous nature of curriculum offerings, ordinary and generalized least squares and profit analysis were used for statistical estimation.

ANDREW G. CUTHBERTSON, Ph.D. North Carolina State 1971. Occupational training in the trucking industry.

The purpose of this study was to analyze the growth of occupational training via formal schooling for long-distance drivers in the trucking industry. Reduced returns to firms offering on-the-job training and public subsidization of schooling may account for the recent rapid expansion in the number of schools and students trained. Estimated rates of return to students' investing in schooling (based on a survey of graduates from one school) were relatively high.

JOHN P. DAVID, Ph.D. West Virginia 1972. Earnings, health, safety, and welfare of bituminous coal workers since the encouragement of mechanization by the United Mine Workers.

ROBERT T. DEANE, Ph.D. California (Los Angeles) 1971. Simulating an econometric model of the market for nurses.

A large-scale model of the U.S. registered nurse market was specified, estimated using 1947-66 data, and simulated from 1947 to 1976. By selectively altering policy variables, sensitivity tests implied that Medicare's impact on nurses was small while nurse training subsidies did not increase nurse graduations. In fact, these subsidies, if effective, would have caused serious unemployment and depressed wages. Suggestions for continued use of the model and for model improvement are made.

DENNIS N. DE TRAY, Ph.D. Chicago 1972. The substitution between quantity and quality of children in the household.

Theoretical and empirical investigation of determinants of desired family size. Child quality and numbers of children are substitutes in production of household commodity "child services." Derived demand equations

are developed from theory for numbers of children and quality per child and estimated using regression techniques and data from 1960 cross-sectional sample of U.S. counties. Shortcomings of data preclude firm conclusions, but in general empirical results support theoretical formulation.

DAVID A. DODGE, Ph.D. Princeton 1972. The structure of earnings of Canadian accountants, engineers, and scientists, and the implications for returns to investment in university education.

The dissertation examines the net effect of university training on annual earnings of Canadian accountants, engineers, and physical scientists. Annual earnings are estimated by *OLS* from a single equation model containing education, experience, work function, specialty, sector, region, and family background variables. Estimated social net present values of returns to investment in education are negative at discount rates exceeding 5 percent for graduate training within the three professions, although errors of estimate are large.

CHRISTOPHER R. S. DOUGHERTY, Ph.D. Harvard 1972. Labor substitution and its implications for educational planning.

Parts 1 and 2 (published as "Substitution and the Structure of the Labor Force," *Econ. J.*, Mar. 1972, and "Estimates of Labor Aggregation Functions," *J. Polit. Econ.*, Nov./Dec. 1972, respectively) use international census data and U.S. state data to examine and measure substitution between different types of labor in production. Part 3 (published as "Optimal Allocation of Investment in Education" in H. B. Chenery, ed., *Studies in Development Planning*, Cambridge, Mass. 1971) outlines the implications for educational planning.

THOMAS F. DUSTON, Ph.D. Brown 1972. Nonpecuniary factors in the occupational choice: A theoretical extension of the human capital model with applications to nursing.

This thesis contains three major sections: the development and use of a technique for estimating the monetary value of nursing nonpecuniary earnings; a nursing supply function derived from an aggregation of explicit distributions of individual preferences; and a test of occupational preference stability as a function of training specificity.

RANDY D. ELKIN, Ph.D. Iowa State 1971. An evaluation of benefit-cost analysis as a tool for manpower decision making.

STEPHEN R. ENGLEMAN, Ph.D. California (Berkeley) 1971. An economic analysis of the Job Corps.

THOMAS J. ESPENSHADE, Ph.D. Princeton 1972. The cost of children in urban United States.

The role of the cost of children in the development of socioeconomic theories of fertility is first reviewed followed by a survey of existing methods of estimating the direct (money expenditure) costs of children. A new method of estimation is then elaborated and applied to

the 1960-61 Bureau of Labor Statistics' *Consumer Expenditure Survey*. Costs are estimated for each child in one, two, and three-child families.

VINCENT J. EVANS, Ph.D. Duke 1972. An analysis of the demand for professors for the State System of Higher Education of North Carolina: 1970-2000.

GREGORY B. FAWCETT, Ph.D. Iowa State 1971. An economic-attitude model for career choice in medicine.

For a maximizing model of career choice in medicine, an attitude-based occupational utility function was postulated to exhibit ellipsoidal contours in five dimensional attitude space. Perceptions of career choice alternatives were reflected as bounded distributional regions in that space, distinguishable by discriminant analysis. The maximizing scheme required that a career be selected whose discriminant projection was tangent to the greatest indifference hyperellipsoid. Subsequent empirical analysis of over 3,000 interns revealed the model's potential in predicting an intern's choice of type of career, type of medical practice, and area of specialization.

LILA J. M. FLORY, Ph.D. Iowa 1972. The stability of the Phillips curve wage rate growth-unemployment tradeoff.

Attention is focused on the controversial long-term inflationary impact of monetary or fiscal policies to decrease unemployment. The theoretical discussion demonstrates that accelerating inflation follows only if policymakers are attempting "over-full employment," but that labor market rigidities may result in substantial periods of unemployment. However, changes in the distribution of unemployment may cause an aggregate Phillips curve to shift. Estimated two-digit industry wage growth-unemployment relationships did not support the accelerationist hypothesis.

MARK S. FREELAND, Ph.D. Wisconsin (Madison) 1972. Determinants of the incidence of work limitations associated with chronic illness and impairments.

ALAN N. FREIDEN, Ph.D. Chicago 1972. A model of marriage and fertility.

This study is a model of population growth containing an examination of marriage based on a theory of the marriage market. Three factors are isolated (the sex ratio, the returns to marriage, and the cost of divorce) which affect reproduction rates through their influence on the prevalence of legal marriage. It is shown that marriage is a purposive decision and that the fertility of younger women is strongly influenced by factors in the marriage market.

VINCENTE G. GALBIS, Ph.D. Wisconsin (Madison) 1972. A contribution to the theory of labor migration and interregional differentials.

WENDY L. GRAMM, Ph.D. Northwestern 1971. A model of the household supply of labor over the life cycle: The labor supply decision of married school teachers.

A utility-maximizing model of household behavior is developed in order to explain the labor supply behavior of married women, especially how the wife's labor supply behavior changes over the life cycle in response to the aging of her children. Empirical work consists of tests of implications of the model based on data collected in a survey of over 400 married women qualified to teach in the northern suburbs of Chicago.

MORLEY GUNDERSON, Ph.D. Wisconsin (Madison) 1972. Determinants of individual success in on-the-job training: An econometric study.

WILLIAM J. HALEY, Ph.D. North Carolina State 1971. Human capital accumulation over the life cycle.

Most of the research on human capital has been the calculation of internal rates of return to specific investments. This thesis develops a more general framework of analysis. A model was developed by assuming that the individual invested in himself with the objective of maximizing the present value of his lifetime earnings. Several functions were derived in parametric form including the path of capital accumulation and the earnings path over the life cycle. The model is simulated for various parameter values to determine how the length of time spent specializing behaves as the parameters change.

FREDERICK J. HEBEIN, Ph.D. Southern Methodist 1972. Investment in human capital: Growth of earnings and rate of return.

A recently developed non-linear model is used to predict annual earnings for groups of individuals with similar characteristics. In this dissertation, the effect of respecifying the model's rate of return parameter from a rental rate to an internal rate is investigated, and found to be insignificant. The present research estimates time-series parameters as well as cross-sectional parameters. One of the findings is that linear autonomous growth decreases as educational attainment increases.

DAVID L. HORNER, Ph.D. Wisconsin (Madison) 1972. The impact of negative taxes on the work effort and wage rates of low-income household members.

The analysis includes a critique of estimates of labor supply parameters obtained from single period cross-sections; an analysis of the work effort response of low income household members to alternative levels of negative taxation; and an analysis of the sources of differences in the wage rate changes experienced by household members subjected to alternative subsidies and tax rates. The data used is from the first year of the Graduated Work Incentive Experiment.

JONATHAN R. KESSELMAN, Ph.D. Massachusetts Institute of Technology. The impact of fiscal redistributive policies on the supply of labor: Five essays in economic theory and program design.

EDWARD J. D. KETCHUM, Ph.D. Princeton 1972. The short-run demand for labor in Canadian manufacturing industries.

This dissertation develops time-series models of the demand for employment and average hours paid-for. A measure of excess labor is developed by making the distinction between hours paid-for and hours worked suggested by Fair, and both employment and hours paid-for are treated as quasi-fixed factors. Several hypotheses are then tested using seasonally unadjusted monthly data for twenty-seven industries and industry groups of the Canadian manufacturing sector.

HAN KIM, Ph.D. Stanford 1971. On the role of human capital in optimal economic growth.

CLAY B. KING, Ph.D. Washington State 1971. Some aspects of collective action, earnings, fringe benefits, and hours of work among senior high school teachers in first-class districts in Washington State, 1960-70. This study is concerned primarily with an analysis of earnings, work hours, and personnel policies of senior high school teachers in the state of Washington.

CHARLES B. KNAPP, Ph.D. Wisconsin (Madison) 1972. A human capital approach to the burden of the military draft.

ARLEEN LEIBOWITZ, Ph.D. Columbia 1972. Women's allocation of time to market and nonmarket activities: Differences by education.

That more educated women are more deterred from work in the labor market by the presence of young children is shown with 1960 census data. Using time budget data, differences in lifetime labor supply patterns by education are traced to differences in time allocation to child care and other household activities. Specifically, more educated women spend more time in child care, although they have fewer children.

ERNE S. LIGHTMAN, Ph.D. California (Berkeley) 1972. The economics of military manpower supply in Canada.

CYNTHIA B. LLOYD, Ph.D. Columbia 1972. The effect of child subsidies on fertility: An international study.

The first part of this study analyzes and tests empirically the effects of various kinds of child subsidies (particularly family allowances) on fertility using international cross-section and time-series data. Although some tests provide weak evidence of their positive effect on fertility, the statistical results were disappointing. The second part reuses the international cross-section data to show the extent to which population considerations influence governments' decisions about the design of income maintenance programs.

MARK A. LUTZ, Ph.D. California (Berkeley) 1972. The equilibrium industrial wage structure: An analysis in terms of wage theory.

JAY R. LYMAN, Ph.D. California (Davis) 1972. Comparative costs of manpower education: A methodological study.

ROBERT A. McMILLAN, Ph.D. California (Berkeley) 1972. Dual labor markets: The case of the urban ghetto Negro male.

MIROSLAV MACURA, Ph.D. Princeton 1972. Estimates of the completeness of registration of births and infant deaths in Yugoslavia and its main provinces from the late 1940's to 1961.

This study presents estimates of annual underregistration of births and infant deaths and true numbers of these events for Yugoslavia and main provinces from late 1940's to 1961. Underregistration estimates of the two types are obtained from vital registration and census data by utilizing two estimation procedures developed in the study. Either procedure consists of estimating underregistration in a pair of cohorts and of converting the cohort rates into period rates of underregistration.

EVELYN MIAO, Ph.D. Wisconsin (Madison) 1972. The structure and performance of the proprietary institutions of higher education in the Philippines.

ROBERT G. MOGULL, Ph.D. West Virginia 1970. Discrimination in the labor market.

CELIA A. MORGAN, Ph.D. Houston 1971. The geographic mobility of labor: An investigation of the role of wages and unemployment rules in the migration process.

JAMES J. MORRIS, JR., Ph.D. California (Berkeley) 1971. Some aspects of the theory of investment in education.

COLLETTE H. MOSER, Ph.D. Wisconsin (Madison) 1971. An evaluation of area skill surveys as a basis for manpower policies.

VICTOR D. NORMAN, Ph.D. Massachusetts Institute of Technology. Education, learning, and productivity.

RONALD L. OAXACA, Ph.D. Princeton 1971. Male-female wage differentials in urban labor markets.

This study estimates what portion of the male-female wage differential can be attributed to the effects of discrimination. Separate wage equations, based on cross-section data taken from the 1967 *Survey of Economic Opportunity*, are estimated for each race-sex group. The results show that sex discrimination accounts for most of the observed wage differential and that such discrimination is less in government employment and in large urban areas.

JOHN L. PALMER, Ph.D. Stanford 1971. Inflation, unemployment, and poverty.

CARL D. PARKER, Ph.D. Oklahoma State 1971. The determinants of hours of work of low-income family heads: A statistical analysis.

The purpose of this dissertation is to estimate the relation between hours worked and sources and levels

of income, as well as hourly wages, controlling for the effects of other market and personal factors. The study utilizes specified groups selected from the 1967 *Survey of Economic Opportunity* data. The primary focus is on those family units who qualify for benefits under the provision of the proposed Family Assistance Plan.

GARY M. PICKERSGILL, Ph.D. Washington 1971. Internal migration in Italy, 1951-61.

LOUIS F. PISCIOTTOLI, Ph.D. Duke 1972. Allocational and distributional effects of conscription.

BETTE H. POLKINGHORN, Ph.D. California (Davis) 1972. The British incomes policy 1964-70: Its impact on the low paid worker.

THOMAS M. POWER, Ph.D. Princeton 1972. Elasticities of substitution between different types of labor: Theoretical analysis and empirical examples.

Six different definitions of the elasticity of substitution when there are more than two factor inputs are compared and their usefulness discussed. Several models for estimating such elasticities are compared. Problems due to differences in labor quality and industry product mix are discussed. The models are applied to Census data with poor results.

TIMOTHY W. PYRON, Ph.D. Louisiana State 1971. Labor union membership: An empirical evaluation of some of the proposed explanations of membership growth. Shift-share and correlation analyses of sectoral changes in employment and union membership for 1958-68 are utilized to assess the claim that the more rapid growth of employment in less unionized categories of employment (specifically workers in the less unionized sectors, nonproduction workers, and workers in the South) is a significant hinderance to union membership growth.

SAMUEL A. REA, Ph.D. Harvard 1972. The supply of labor and the incentive effects of income maintenance programs.

The theory of labor supply is analyzed with special attention given to the relation between labor supply and unemployment. Supply functions for those age 25 and over are estimated, and the supply response to eleven different negative income tax plans is simulated. A \$2400 guarantee for a family of four with a 50 percent tax rate is estimated to reduce hours supplied by the recipients by 12 percent and decrease the number in the labor force by 21 percent.

STEPHEN M. RENAS, Ph.D. Georgia State 1971. An economic analysis of academic dropouts.

A model is developed in which an individual selects that educational investment option which enables him to maximize utility, subject to constraints imposed in the capital market. We estimate by simulation the degree to which a person's willingness to defer gratification and ability to finance his education may affect his decision to obtain an education. We demonstrate that

the decision to remain in school is mildly affected by one's success in obtaining financial assistance and willingness to sacrifice present for future consumption.

FREDRICKA P. SANTOS, Ph.D. Columbia 1972. Some economic determinants of marital status.

This study investigates whether complementarity between husbands and wives in the production of home and market goods is an important factor determining marital status. In order to analyze the relationship between marital status and female market potential (holding family income and other relevant factors constant), multiple regression techniques were used. U.S. Census data for the years 1950 and 1960 facilitated the empirical analysis. The results from running weighted least squares across forty-eight states supports the hypothesis. That, is, complementarity between husbands and wives is a significant factor determining marital status.

GERALD E. SCHLUTER, Ph.D. Iowa State 1971. An estimation of agricultural employment through an input-output study.

An interregional input-output model emphasizing agriculture was used to study the employment structure of the agri-business complex and to estimate the magnitude of 1964 employment requirements for food and natural fiber delivered to final demand. Projections of this measure of agricultural employment were made to 1980. A survey was made of the implications for rural communities. Educational planning was projected within areas of lower agriculture employment requirements.

LARRY S. SCHROEDER, Ph.D. Wisconsin (Madison) 1972. Occupational and geographical mobility within Wisconsin, 1946-60: An economic analysis.

DAVID SHAPIRO, Ph.D. Princeton 1972. Three aspects of the economics of education in Alberta.

The study examines three aspects of the economics of education in Alberta: the demand for teachers; the mobility of teachers; and the relationship between school district size and the level of costs/expenditures on education. The individual school district is the focus of analysis. An analogy is drawn between the school district and the firm, and implications for school district behavior are developed from the theory of the firm.

LAWRENCE SLIFMAN, Ph.D. Washington (St. Louis) 1971. Occupational mobility of low income workers.

This dissertation is concerned with assessing the impact of changes in the aggregate labor supply-demand balance on the upward occupational mobility of lower income workers. A labor market model emphasizing the probabilistic nature of labor market decisions is examined, and the results suggest that generally upward mobility is stimulated by the timing of life cycle job changes in response to tightened labor markets.

GARY H. STERN, Ph.D. Rice 1972. Price expectations, inflation, and the labor market.

This research is concerned with the neoclassical expectations view of the inflation-unemployment relation. There were three distinct purposes to the empirical work. In particular, the role of price expectations in the wage determination process was a major issue and, secondly, appropriate representation of the excess demand for labor was explored in some detail. Finally, tests of the stability of the Phillips relation within the 1952-58 sample period were performed.

JOHN M. SWINT, Ph.D. Rice 1972. A multiparty collective bargaining model and its implications.

This study develops a multiparty collective bargaining model to investigate both the causes of strikes in the private nonagricultural sector and the factors determining their duration. The model's active negotiating participants, union leaders and management, maximize their joint utility subject to both economic and political constraints which are imposed upon them by their respective constituents, union members and stockholders. Two sets of hypotheses are derived from the model and then tested using conventional econometric techniques.

RICHARD TOIKKA, Ph.D. Wisconsin (Madison) 1972. Supply responses of the unemployed: A probability model of reemployment.

CARMEL J. ULLMAN, Ph.D. Columbia 1972. The professional in the labor force: An econometric study of the growth of professional and technical occupations in the United States, 1880-1960.

Aggregate demand for professional and technical personnel relative to other workers is derived from a production function specifying two human capital factors corresponding to the two kinds of labor. Aggregate supply is derived from human capital theory. Estimated demand and supply functions are found stable throughout the period. Biased technological change is shown unimportant for explaining professionalization; changing occupational distribution represents factor substitution induced by increasing relative supply of professionals.

JOHN M. VOLPE, Ph.D. New York 1972. The effect of the multinational corporation on domestic labor in manufacturing.

This study, in part, analyzes world-wide labor cost differentials, outlines parent firm and subsidiary capital outlays and subsidiary sales by industry, develops techniques to determine whether capital outlays at home and abroad are in industries characterized as human capital, physical capital, or labor extensive, and surveys some significant trends in employment within the service-oriented industries.

STEPHEN A. WANDNER, Ph.D. Indiana 1972. Racial patterns of employment in Indianapolis: The implications for fair employment practices policy.

This dissertation examines aspects of the economics of employment discrimination against Negro males in the Indianapolis Standard Metropolitan Statistical Area (SMSA). Measures of employment discrimination within the SMSA are constructed from statistical data

on Negro and Caucasian male employment from the Equal Employment Opportunity Commission's annual employment census. These measures are then used to evaluate the effectiveness of the policies of the fair employment practice commissions operating in Indianapolis in identifying and eliminating employment discrimination.

RICHARD F. WERTHEIMER, Ph.D. Maryland 1971. The monetary rewards of migration within the United States.

The purpose of this study is to estimate the economic benefits of migration to the migrants themselves. Hypotheses relating the size of earnings gains to the education, race, and sex of the migrants are tested using multiple regression analysis to obtain estimates of the annual earnings differentials attributable to migration. The findings indicate there are substantial gains to be had by moving out of the South and by moving from rural areas into cities.

DONALD R. WINKLER, Ph.D. California (Berkeley) 1972. The production of human capital: A study of minority achievement.

### Welfare Programs; Consumer Economics; Urban and Regional Economics

JAMES W. ADAMS, D.B.A. Georgia State 1971. A study of the effect of government aid and other factors on the economic development of three selected counties in Georgia.

This is a study of the dominant factors influencing economic development of underdeveloped rural areas with particular emphasis upon the effect of government aid as an influencing variable. Economic changes and prevailing socioeconomic influences are traced from 1960 through 1967 in three selected rural counties in Georgia, employing a case study approach. It is concluded that geographical juxtaposition in relation to developed areas constituted the dominant influencing variable. Government aid and other factors were not dominant variables.

ROGER S. AHLBRANDT, JR., Ph.D. Washington 1972. Efficient output of a quasi-public good: Fire services.

TERRY L. ANDERSON, Ph.D. Washington 1972. The economic growth of seventeenth century New England: A measurement of regional income.

NICHOLAS A. BARR, Ph.D. California (Berkeley) 1971. Public assistance and family behavior in the urban United States.

KURT R. BAYER, Ph.D. Maryland 1971. A social indicator of the cost of being Black.

This study sets forth a framework for analyzing the negative effects of racial discrimination (past and

present) on the well-being of Black urban families. By converting these effects occurring in the areas of public education, the housing market and higher exposure to crime into income equivalents, and adding these to the income differential between Black and white families, the actual well-being situation of Blacks relative to whites is exhibited more clearly than through the use of income differentials alone.

DEVINDER K. BHATIA, Ph.D. Pennsylvania State 1972. Income distributional implications of regional expenditure programs using the interindustry model.

MICHAEL K. BLOCK, Ph.D. Stanford 1971. An economic analysis of theft with special emphasis on household decisions under uncertainty.

In this dissertation two sets of theft related decisions are analyzed. Both the decision by the offender to commit theft and the decisions by victims to modify the adverse effects of theft are considered. Because of the nature of the theft and security decisions, the element of uncertainty is central to the analysis. The results in both decision problems are distribution free and are generalizations of the existing literature in this field.

ROBERT A. BOHM, Ph.D. Washington (St. Louis) 1971.

The determinants of nonwhite male employment growth and migration behavior in large U.S. cities.

This study attempts to determine for nonwhite males the effect of migration on employment growth, and the extent that migration responds to economic influences. Estimates of the elasticity of nonwhite male employment change with respect to migration indicate that nonwhite male migrants have difficulty finding jobs in urban labor markets. However, nonwhite male migrants respond strongly to the rate of change of employment with respect to labor force and interurban differences in income.

BRUCE R. BOLNICK, Ph.D. Yale 1972. Charity and the free rider: Consideration for the theory of public goods.

ELIYAHU BORUKHOV, Ph.D. Johns Hopkins 1972. City, size, land use, and transportation costs.

PHILLIP D. BROOKS, Ph.D. Kansas State 1971. An appraisal of the primary and secondary data approaches for implementing single region input-output models.

DOUGLAS M. BROWN, Ph.D. West Virginia 1970. Productive capacity and economic growth in West Virginia.

DENNIS R. CAPOZZA, Ph.D. Johns Hopkins 1972. Transportation and the urban economy.

JOHN J. CASEY, Ph.D. Georgetown 1972. Economics of narcotics addiction.

JEFFREY I. CHAPMAN, Ph.D. California (Berkeley) 1971. A model of crime and police output.

This dissertation develops a four-equation simultaneously estimated model describing the causes of crime, police output with respect to crime, and society's demand for police, taking crime into account. It was found that relative illegal wages have little influence on crime, arrest rates cannot be said to definitely retard crime, police labor is positively related to arrest rates, and property crimes are more important than other types of felonies in increasing the demand for police.

SANDRA CHRISTENSEN, Ph.D. Wisconsin (Madison) 1972. Income maintenance and the labor supply.

KWANG-WEN CHU, Ph.D. California (Los Angeles).

Consumption patterns among different age groups: An econometric study of family budgets.

The primary purpose is to use cross-section data to compare the allocation of consumption expenditures in the budgets of households headed by persons of different ages. Some working hypotheses were formulated within the framework of neoclassical theory of consumer demand. A model of family budget allocation and a generalized expenditure function of households were constructed and used in testing the hypotheses against the data of the *Survey of Consumer Expenditures*, 1960-61.

RONALD L. COCCARI, Ph.D. West Virginia 1971. A regional linear programming model of the West Virginia economy.

GAVIN L. COLLINS, Ph.D. Iowa 1972. A normative framework for determining the rate of discount and for making capital expenditure decisions within the health care industry.

A two-part study presenting a synthesis of the bits and pieces of economic theory relevant to a choice of discount rates for both profit-seeking firms throughout the economy and an empirical estimate of the risk-adjusted discount rate for hospitals in Iowa at the end of 1968.

ERNEST F. COMBS, Ph.D. Washington 1971. The economics of investment programs resulting in the prevention, early detection, and early treatment of chronic illness.

ROBERT H. EDELSTEIN, Ph.D. Harvard 1972. Essays on capital allocation problems and urban analysis.

This thesis consists of three interrelated essays, each focusing upon one aspect of contemporary social-economic problems. There are four intertwining elements common to all essays: each essay explores some aspect of urban analysis. Essay One discusses the workings of the urban property insurance market. Essay Two develops and empirically tests a theory for commercial bank lending policies to minority enterprises. Essay Three examines the economic issues behind Black economic development.

JOHN D. ESHELMAN, Ph.D. Washington 1971. Resource allocation in medical research.

RAYMOND L. FALES, Ph.D. Northwestern 1971. Loca-

tion theory and the spatial structure of the nineteenth century city.

A study of urban location theory which incorporates von Thunen type and Weberian location theory to explain the spatial location of industry and households of 1873 Chicago.

JAMES M. FITZMAURICE, Ph.D. Maryland 1972. The demand for hospital services: An econometric study of Maryland counties.

A model is developed relating a county's consumption of hospital services to economic characteristics, socio-demographic characteristics, and the availability of alternative sources for medical care. The data consists of a cross-section of Maryland's 24 counties, their 44 short-term general, nonprofit hospitals for the period 1965-68. Hypotheses testing the influences of Medicaid, Medicare, price, income, the availability of physicians, hospital beds, nursing home beds, age, and other factors were undertaken with significant results.

JAMES L. FREUND, Ph.D. Washington (St. Louis) 1971. Differential short period changes in earned income in manufacturing among urban areas.

A labor market model for urban areas is developed to explain short period earnings changes. Both national and local economic forces are posited to be important. Earnings changes in manufacturing among urban areas are best explained by local labor demand in short periods, with national forces being relatively more important over longer periods. The study puts past results into perspective and has implications for the temporal applicability of theories of urban economies.

CHARLES J. GALLAGHER, Ph.D. West Virginia 1971. A study in the use of national input-output data in regional input-output analysis.

MORRIS B. GOLDMAN, Ph.D. California (Los Angeles) 1971. The economics of prepaid medical care: A capital theoretical approach recognizing uncertainty.

This study examines the relationship between physician remuneration methods and the amount of medical care consumed by patients. The model, based upon the costliness of health care information, intertemporal decision process, and differing consumer risk preferences, implies that in long-run equilibrium, patients choosing prepaid practitioners will consume less medical care than fee-for-service patients. The viability of this association crucially depends on consumer preferences and not physician incentives per se. An empirical literature review provides supporting evidence.

ORVILLE F. GRIMES, JR., Ph.D. Chicago 1971. An economic analysis of recreational land under urban influence.

Recreational land price, lot size, and housing density were determined in a model of private access recreation near metropolitan areas. The spatial pattern of demand for land emanating from more than one urban center was analyzed. Beach accessibility, exclusiveness of neighborhoods, water quality, and other amenities entered the model through their effects on land price.

Empirical results, obtained from a large sample of Lake Michigan properties, substantially confirmed the predictions of the theory.

ANDREW M. HAMER, Ph.D. Harvard 1972. The comparative costs of location of manufacturing firms in urban areas: A Boston case study.

Profit maximizing location decisions for a wide range of urban manufacturing firms collapse into cost minimization with few opportunities for alteration of land use intensity or the composition of labor by skill grades. Land and property tax costs variations therefore play a commanding role in location. These variations, incorporated into hypothetical gross rentals for model firms in various industries, favor suburban locations. These differentials have a serious impact on hypothetical net operating profits of urban firms.

BRUCE W. HAMILTON, Ph.D. Princeton 1972. The impact of zoning and property taxes on urban structure and housing markets.

A model of an urban area containing many municipalities is constructed. Each municipality provides its residents with some level of public service, financed by a property tax, and may require its residents, through zoning, to consume some minimum amount of property. Under some restrictive assumptions, this model yields a Pareto optimal level of public service provision. And the property tax, acting as a price, yields no deadweight loss. Empirical tests are reported.

RAYMOND W. HAMILTON, Ph.D. Maryland 1971. The public housing program in the United States: An analysis and evaluation.

This study examines economic aspects of the public housing program. Average aggregate monthly subsidy for total inventory approximately \$80; for newly developed units \$150. Unique features of public housing enable it alone among federal programs to provide decent shelter for the lowest income families. It does so, without reducing the satisfaction of other families, more efficiently than cash grants. It provides a direct link between supply and demand, and, hence, precludes adverse market effects.

ROBERT P. HAMRIN, Ph.D. Wisconsin (Madison) 1972.

Performance contracting in education: An economic analysis of the 1970-71 Office of Economic Opportunity Experiment.

The purpose of this study is to analyze three economic aspects of the experiment: the experimental design specifications for selection of the participants; the payment structure and its implicit behavioral price incentives; and the relationship of the instructional input variables to the achievement test results. The analysis revealed that the specifications contained many experimental design and distributional problems, and the experiment's evaluations were incomplete by not modelling the large differences in instructional inputs.

THOMAS W. HELMINIAK, Ph.D. Wisconsin (Madison) 1972. The sugar-bananas shift on St. Lucia, West Indies: Bilharzia and malaria disease causal linkages.

- J. VERNON HENDERSON, Ph.D. Chicago 1972. The types and sizes of cities: A general equilibrium model.
- JOHN HOLAHAN, Ph.D. Georgetown 1971. Benefit-cost analysis of programs in the criminal justice system.
- DENNIS A. JOHNSON, Ph.D. Iowa 1972. Poverty in the United States: Its meaning and measurement.  
A definition of poverty which takes account of non-money income is developed. This definition is incorporated into a model which allows empirical estimates of the effect of some nonmoney incomes on the poverty income gap to be made. Consideration of nonmoney income importantly affects the size of the poverty income gap and the movement of poverty through time, and also challenges the belief that increasing money income is the most efficient way of reducing poverty.
- PETER KARPOFF, Ph.D. Wisconsin (Madison) 1971. A proposal for reforming nursing home reimbursements under Medicaid.
- ROBERT A. KELLY, Ph.D. Georgetown 1971. Housing goals and the housing mix.
- UNG SOO KIM, Ph.D. CATHOLIC. Measuring and analyzing the impact of employment generation benefits of public water resource development project in Appalachia.
- A. THOMAS KING, Ph.D. Yale 1972. Land values and the demand for housing: A micro-economic study.
- WILLIAM R. KRAUSE, Ph.D. Maryland 1971. Evaluating and forecasting progress in racial integration.
- LILY KUO LAI, Ph.D. Wisconsin (Madison) 1972. The estimation of effects of expected family income and socioeconomic variables on the U.S. household consumption of food commodity groups.
- CHARLES E. LEITTLE, Ph.D. Arkansas 1972. Adjustment of a community from an exhaustible resource base to other economic alternatives: A case history of Joplin, Missouri.  
A case study to test the hypothesis that when a region is built on an exhaustible resource, the people who control it will also control the development of the region in that in the long-run, they do not invest sufficiently in long-term projects such as social overhead capital. The transition of the community to other economic alternatives was also included.
- DANIEL S. LEVINE, Ph.D. Northwestern 1971. Economic development in Appalachia.  
Factor analysis is used to investigate the interrelationships of economic, social, and political behavior in the long and short-run development processes of 574 counties in Appalachia and its immediate environs. Sixty variables are used. The primary result is to indicate that differences in economic behavior and institutions explain the bulk of income variation. However, the association patterns of these variables are more complex than is generally allowed for in regional development analysis.
- WILLIAM P. LILES, Ph.D. South Carolina (1972). The effect of racial discrimination on earnings and employment: Additional evidence.
- ROGER W. LIND, Ph.D. Maryland 1971. Determinants of local public expenditures: A study of Rhode Island's thirty-nine cities and towns.
- CHARLES D. LINER, Ph.D. Washington (St. Louis) 1972. A study of family medical care expenditures.  
This study examines the effects of family income, family size and composition, race, education, occupation, and health insurance coverage on family medical care expenditures. A theoretical framework for analyzing medical care demand by families is presented, and the hypotheses are tested using multivariate regression techniques on cross-sectional data from a survey of 645 household in urban communities of the St. Louis area.
- BEN-CHIEH LIU, Ph.D. Washington (St. Louis) 1971. Regional growth and local government finance: An econometric study.  
This study attempts to explore the interdependent relationship between regional employment growth and local government finance. In a simultaneous equations model, net migration rate, employment growth, and changes in local government expenditures per capita and the average tax rates were hypothesized to be determined endogenously. Empirical tests utilizing data collected from the seventy-five largest SMSAs in the United States, 1960 to 1967, were conducted and the newly developed hypotheses in this paper were accepted.
- JOHN F. McDONALD, Ph.D. Yale 1971. Essays on the pattern of home ownership and retail trade: A case study of Detroit.
- PETER D. MACHLIS, Ph.D. Rutgers 1971. The distributional effects of public higher education in New York City.  
Distributional effects are determined by distributing the current expenditure of New York City and New York State for public higher education according to the percentage distribution of city and state taxes and the percentage income distribution of students' families. This provides an estimate of the taxes paid by families in each income class to support public higher education, and the direct monetary benefits received by families in each income class from public higher education.
- OLIVER L. E. MBATIA, Ph.D. Oregon State. Economic impact of the 1964 Fair Employment Practices Act and subsequent executive orders on Black Americans.  
The objective of this thesis is to examine the effects of Fair Employment Practices (FEP), 1964, on the relative economic status of Blacks. A theoretical model shows that racist employers see two different market demand curves for Black and non-Black workers. FEP laws are intended to shift Black demands to the right,

equalizing that of non-Blacks. Statistical tests shows *FEP* coefficients to possess negative sign. Therefore, *FEP* laws could improve the economic status of Black Americans.

M. AMATA MILLER, Ph.D. California (Berkeley) 1971. Regional differences in real wages: United States, 1860-80.

RALPH C. MOOR, JR., Ph.D., Georgia State 1971. The provision of education in economic space.

KENT D. NASH, Ph.D. North Carolina State 1972. The demand for hospital construction and other expenditures for medical care.

A standard demand model was estimated for a set of cross-section/time-series Hill-Burton data. Varying construction subsidies caused price variation. Demand for new hospital beds was relatively inelastic but subsidies lowered price enough so that about one-third of all the beds built in the period were program induced. The Hill-Burton program in North Carolina changed construction timing but not identity for communities that would otherwise have built new facilities.

GARY R. NELSON, Ph.D. Rice 1972. An econometric model of urban bus transit operations.

This thesis develops and estimates a simultaneous-equations model of the urban bus transit market. The econometric model consists of a demand equation and two equations derived from a general model of supply behavior. The parameters of the model are estimated from cross-section data in 1960 and 1968 on about fifty transit monopolies. The estimates are used to test a set of hypotheses about the economic behavior of transit firms and regulators.

THAE SOO PARK, Ph.D. Wisconsin (Madison) 1972. Old-age assistance expenditure determination: An empirical study.

RONALD L. PROMBOIN, Ph.D. Stanford 1971. Regional impact of the U.S. interstate highway system: A macro-economic approach.

This dissertation was an investigation of the macro-economic impact of the interstate highway system on states. An attempt was made to separate two effects, short-run employment creation and long-run capital good. Because of the nature of the models and data used, results were chiefly qualitative and suggestive. Employment creation effects were more significant, particularly in the West, while the longer-term productivity of the system appeared (weakly) significant chiefly in the Southeast.

PHILIP ROBINS, Ph.D. Wisconsin (Madison) 1972. A theory and test of housing market behavior.

WILLIAM D. ROHLF, JR., Ph.D. Kansas State 1972. An analysis of urban size as a factor influencing the growth of manufacturing industries.

The influence of urban size on the recent growth of manufacturing industries was investigated using a

cross-sectional multiple-regression model. Productivity growth, wage rates, and distance were among the other variables included in the model of urban manufacturing growth. For the sixty-six industries analyzed, only five showed a significant determinant between growth rates and urban size and these all had negative signs. The study interval was 1963 to 1967.

JOHN D. F. ROWLATT, Ph.D. Princeton 1971. Welfare and the incentive to work: The Alberta case.

The costs of altering the regulations of a welfare system depend upon the labor supply response of low-income people to changes in the after-tax wage rate and the guarantee level. This research analyzes samples of Alberta welfare recipients and finds that their duration of assistance is consistent with the predictions from a static labor supply model. In addition, the actual welfare tax rate is estimated and is significantly less than the statutory rate.

JOHN G. RUFF, Ph.D. Kansas State 1971. Economic analysis of the structural change of an agricultural region: The case of the Kansas economy, 1954-67.

GORDON A. SAUSSY, Ph.D. Yale 1972. Modeling the location of basic manufacturing in the New Orleans metropolitan area.

MALINDA SCHAILL, Ph.D. Northwestern 1972. Differential growth of employment in Illinois cities.

This dissertation focuses on the nature of the city as a location for production, and offers a model describing the growth of employment in cities based on the urban amenities of each place. The theoretical model incorporates site advantages into the efficiency of the firm's production function, as well as incorporating the price of land and the property tax into the firm's cost structure. The model is tested on a sample of seventy-two Illinois cities.

JOHN A. SCHOFIELD, Ph.D. Simon Fraser 1972. Cost-benefit analysis and the policy makers' objective function: The case of British regional policy, 1960-66.

This thesis has two aims. The first is to estimate the tangible efficiency returns to contrasting regional policies (distribution of industry policy and labor migration policy) as they were employed in Britain, 1960-66. Separate analyses are conducted for the sub-periods 1960-63 and 1963-66 and from the distinct points of view of the economy as a whole and the national government. The second aim is to derive from the results relating to the economy viewpoint, implications regarding the relative importance of different objectives of the regional program. Thus the policy makers' estimated objective function (1960-63 and 1963-66) is revealed.

RICHARD E. SCHULER, Ph.D. Brown 1972. The interaction between local government and residential location in an urban area.

The many relationships between local government activity and urban residential location are examined within a theoretical, spatial-equilibrium framework.

Assuming one area-wide political jurisdiction, the existence and properties of an equilibrium are demonstrated for different income distributions. The long-run welfare implications of alternative levels, spatial distributions, and methods of financing public services are then examined. Only those services primarily designed to offset externalities which arise from increasing population density are considered.

MARQUIS R. SEIDEL, Ph.D. Maryland 1971. Effects of real estate taxes on land use changes.

VISHWA SHUKLA, Ph.D. Wisconsin (Madison) 1972. Farm family expenditure functions: An econometric analysis.

MICHAEL A. SPINELLI, Ph.D. West Virginia 1971. A definition of the economic subregions of Appalachia using factor analysis.

EARL O. STEPHENS, Ph.D. Oregon 1972. An analysis of consumer's purchase of automobile installment credit.

LEANNA STIEFEL, Ph.D. Wisconsin (Madison) 1972. Economic exploration of teaching assistantship stipends at large public universities.

WILLIAM J. STULL, Ph.D. Massachusetts Institute of Technology. An essay on externalities, property values, and urban zoning.

H. BRUCE THROCKMORTON, Ph.D. Arkansas 1972. Cost sharing in the extension of municipal water lines in Arkansas.

The hypothesis states that an extension of a municipal water line by a property developer creates greater benefits than costs to the three units (property developer, service user, and city) in total, but that current policies might preclude one of the units from gaining greater benefits than costs. Data collected from several cities were used to establish a procedure, by means of benefit-cost analysis, which would approach an equitable cost-sharing solution.

THOMAS H. TIETENBERG, Ph.D. Wisconsin (Madison) 1972. Pollution control and the price system: A general equilibrium analysis.

ROBERT B. VERNON, Ph.D. Brown 1972. The structure and growth of a von Thünen economy.

This thesis derives the aggregate structure of a von Thünen economy directly from its spatial base. The model is reformulated in terms of intersectoral flows and optimal land uses found for given populations. Short- and long-run equilibriums for competitive and centralized economies are compared. Competitive equilibriums are demonstrated to be neither unique nor optimum. Growth paths and maximum supportable populations are discussed as functions of population growth and the capital accumulation process.

THOMAS T. VERNON, Ph.D. Kansas State 1971. Public expenditures and location as contributing to small city growth.

This study investigates the influence of urban public capital investment on the population growth of small cities. A hypothesis suggesting that the demand for urban public capital arises from local desires for growth is tested. Data on urban capital formation were obtained from changes in bonded debt during a four-decade period. Significant lag relationships between public capital formation and population growth establishes support for the hypothesis. Finally, small city growth is influenced by location.

ROBERT E. WUNDERLE, Ph.D. Cornell 1971. Evaluation of the pilot Food Certificate Program.

This project involved an economic and nutritional evaluation of the Food Certificate Program which was piloted in Chicago and Bibb County, Georgia. Results indicated that the program was functioning primarily as an income maintenance program and was not producing the desired nutritional results.

FREDERICK C. YEAGER, Ph.D. West Virginia 1972. Personal bankruptcy in the United States with special reference to West Virginia.

STEPHEN H. ZELLER, Ph.D. Boston College 1971. The urban firm's production-location decision: Towards a theory of land and property taxation.

The profit-maximization model is based on a Cobb-Douglas production function in land, labor, and capital. Prices are exogenous and the firm locates along a straight line. Solutions are obtained with computer-aided numerical analysis techniques. The long-run comparative-static behavior of the firm is developed and it is seen that the imposition of either a land or a property tax will cause the firm to reduce its equilibrium stock of capital.

JOSEPH A. ZIEGLER, Ph.D. Notre Dame 1971. An analysis of interurban differentials in economic activity.

The purpose of this study is to investigate the nature of the relationship between national and urban business activity and to determine the factors which explain the varying strengths of this relationship. The analysis reveals that although the average cyclical behavior of the cities conforms generally to the national average, there were wide dispersions in every measure of cyclical performance. Generally, the dispersion is explained by differentials in industrial composition, growth, city size, and urban finances.

DENNIS ZIMMERMAN, Ph.D. Washington (St. Louis) 1971. The distribution of public higher education expenditures and finances: The St. Louis-St. Louis County junior college district.

A case study tests the empirical relevance of four potential methodological deficiencies of the literature on the distributional impact of education expenditures: treatment of the sometimes conflicting goals of efficiency

and equity; choice of criteria for evaluating the observed distributional impact; level of aggregation; and the benefit-expenditure relationship. These empirical results, combined with a tax incidence study, indicate a positive relationship between a student's direct net subsidy and family income.

BARBARA S. ZOLOTH, Ph.D. Minnesota 1971. The economic effects of racial discrimination in secondary schools.

This study primarily involves linear estimation of the level of formal schooling chosen by both Black and white individuals as a function of the individual's socioeconomic background, test scores, expenditure characteristics, and racial composition of his high school, and region of residence. The data was drawn from Project Talent, and the results indicate that for both racial groups test scores and socioeconomic variables are positively significant whereas high school characteristics are not.

## ANNOUNCEMENT

### NOTICE TO ALL GRADUATE DEPARTMENTS

The December 1973 issue of the *Review* will carry the seventieth list of doctoral dissertations in political economy in American universities and colleges. The list will specify doctoral degrees conferred during the academic year terminating June 1973. This announcement is an invitation to send us information for the preparation of the list. This announcement supercedes and replaces a letter which was sent annually from the managing editor's office.

The *Review* will publish in its December 1973 issue the names of those who will have been awarded the doctoral degree since June 1972, the titles of their dissertations, and, if possible, a brief (75-word) summary of the dissertation.

By June 30, please send us this information on 3×5 cards, conforming to the style shown below, one card for each individual. Please indicate by a classification number in the right-hand corner the field in which the thesis should be classified. The classification system is that used by the *Journal of Economic Literature* and printed in every issue.

Name: LAST NAME IN CAPS: First Name, Initial _____	JEL Classification No. _____
Institution Granting Degree: _____	
Degree Conferred (Ph.D. or D.B.A.) _____ Year _____	
Dissertation Title: _____	
Summary (75-word maximum, or first 75 words will be printed)	
Summary may be completed on back of this card or on new card which should be stapled to this.	

When degrees in economics are awarded under different names, such as Business Administration, Public Administration, or Industrial Relations, candidates in these fields whose training has been *primarily in economics* should be included.

All items and information should be sent to the Assistant Editor, *American Economic Review*, Box Q, Brown University, Providence, Rhode Island 02912.